Proteomics and protein activity profiling:

An investigation into the salivary proteome and kinase activities in various systems using mass spectrometry

Fiona E. McAllister

The University of Edinburgh

Doctor of Philosophy by Research

June 2010

Declaration

This thesis is submitted in part fulfilment of the requirement for the degree of Doctor of Philosophy at the University of Edinburgh. Unless otherwise stated, this work is my own and has not been previously submitted, in whole or in part, for any degree at this or any other university.

Fiona E. McAllister, June 2010

Abstract

Protein identification and quantitation using mass spectrometry has evolved as the dominant technique for studying the protein complement of a system: cell, tissue or organism. The proteomics of body fluids is a very active research area as there is great potential for protein biomarker discovery; application of such technologies would revolutionise medical practice and treatment. Saliva, through its non intrusive nature of sampling, is an ideal body fluid for disease diagnosis, screening and monitoring. Gingivitis is a gum disease with symptoms including bleeding, swollen, and receding gums. After dental decay, gingivitis is estimated to be the most common disease worldwide, and around 40% of the population in the US are reported to have gingivitis. The end point goal of this project was to identify salivary biomarkers for gingivitis.

This dissertation presents an investigation of: 1) the salivary proteome; 2) developments and applications of a mass spectrometry kinase assay; and 3) salivary biomarkers for gingivitis using proteomics and kinase activities.

The soluble portion of the human salivary proteome (saliva supernatant) has been studied by several research groups but very few proteomic studies have been performed on the insoluble, cellular and bacterial portion of saliva. Presented here, is the first global proteomics study performed on the saliva residue and supernatant from the same test subject. A total of 834 and 1426 proteins were identified in the saliva supernatant and residue, respectively. A global analysis of protein complexes in saliva was also performed and is the first study, to date, of such an analysis. KAYAK ('Kinase ActivitY Assay for Kinome analysis') was further developed for its application on a number of cell types, tissue types, and a variety of organisms. Proof of concept work for in-gel kinase activity/kinase abundance correlation profiling using blue native gels was performed, and experiments using anion exchange chromatographic kinase activity/kinase abundance correlation profiling were performed to identify kinase-substrate pairs. KAYAK applications included the analysis of kinase activities in *Saccharomyces cervisiae, Drosophila*, mouse, and human saliva in which significant kinase activity was detected in the saliva supernatant, a novel finding. Finally, gingivitis was induced in patients, and the saliva samples were analysed using proteomics and kinase activity profiling. Although this work is ongoing, preliminary data indicate that there are increases in various inflammatory proteins, certain bacteria and also in the activity of particular kinases as a result of the induction of gingivitis.

The overall study provided insights into the salivary proteome for both the human and bacterial complement, as well as discovering the presence of significant kinase activity in saliva. In the induced gingivitis study, almost half of all the proteins identified in the residue were from bacteria (1274 bacterial proteins, 198 species identified) and there may be more potential for biomarker discovery for certain diseases in the saliva residue than in the supernatant. A very large overlap was observed between the human proteins in the saliva supernatant and residue, indicating that many of the salivary proteins originate from lysed cells. The origin of the kinase activity in the saliva supernatant is not known but is also proposed to originate predominantly from lysed cells. A range of novel KAYAK applications have been investigated, demonstrating that KAYAK has a wide variety of future uses ranging from target compound evaluation in Pharmaceutical companies to patient testing in the clinic.

Acknowledgements

When it comes to giving credit to people for their help, support and friendship during the past few years of my studies, the list is long and there are very many to whom I am deeply indebted.

First I should like to thank my supervisor Dr. Perdita Barran at The University of Edinburgh and my secondary supervisors, Drs. Juri Rappsilber and Dominic Campopiano and Prof. Peter Tasker, also at Edinburgh, for their interest and support, particularly in the first two years of my PhD. All four gave me the confidence to develop ideas and helped me to follow different lines of enquiry as well as encouraging me to explore different areas as my work progressed. For this I am truly grateful to them.

Complementing the academic support, Drs. Michael Cannon and Gordon Davison at Procter & Gamble (Egham) provided me with useful information and contacts to which I could always refer. I would like to express my gratitude for the help they gave me along the way. I would also like to thank Dr. Jean Wevers (Procter & Gamble, Brussels) for helping me set up this PhD initially. In the background, but no less important, has been the support given to me by Dr. Regine Labeque also of Procter and Gamble (Brussels). Despite her heavy workload she has always been happy to give me practical advice when I needed it and which I always very much appreciated. I cannot thank her enough for the unstinting support she gave me during the whole of my industrial year and throughout my post graduate period. She is an exceptional role model for me.

For his friendship, as well as for giving freely his advice built on a wealth of knowledge of industry and academia, I have a great deal for which to thank Professor Peter Tasker. Not only was he my Masters supervisor during my undergraduate degree, he has also given me his expert advice throughout my PhD. I doubt whether I would have been able to achieve so much without his sustained encouragement.

At this point I feel it appropriate to express my sense of everlasting debt to all the Trustees of the Kennedy Memorial Trust, who put their faith in my ability to take full advantage of spending a year in the Gygi lab as a Kennedy scholar by awarding me one of the 2008-09 scholarships. I cannot speak highly enough of the way the programme is organised to bring together post graduates of different disciplines for a very exciting and challenging year of study and social activities. Annie Mason and, later, Anna Thomas – secretaries to the trustees – both deserve particular mention not only for the example they give of exceptional organisational skills, but also for the warm friendship and personal kindness they show in their dealings with all the Kennedy scholars during their year at Harvard.

Professor Steven Gygi of Harvard Medical School has been my mentor and inspiration for the second half of the time spent on research for my PhD. Working in his lab is both a stimulating and enjoyable adventure; and learning from his wealth of experience and expertise in the field was an opportunity not to be missed. His eager and generous spirit combined with his sense of fun in endeavour provided memories which will always stay with me. I cannot thank him enough for the quiet confidence he has shown in me. His scientific and personal integrity is second to none.

In any research group there are always a great number of individuals happy to give support both practical and moral. In the Barran group I was immediately struck by the notable team spirit evident from the outset, as much in the weekly group meetings I attended as in the many invitations to join in sporting and leisure activities. Stefan Esswein, Martin De Cecco, Hayden Eastwood, Peter Faull, Jason Kalopothakis, Hannah Florance and Wutharath Chin were good friends as well as good colleagues. Likewise I would like to thank all the members of the SIRCAMS facility who were always happy to give me help and from whom I learned a great deal about mass spectrometry:

To be specific there are many individuals in Edinburgh who gave me invaluable practical help in the early stages and I should like to thank very sincerely all the following: Dr. Holger Husi for his help and expertise in proteomics, and for introducing me to gels and phosphoproteomics; Dr. Logan Mackay for very generously and kindly letting me use the HCT and FTICR mass spectrometers; Yana Berezovskaya and Drs. Stefan Weidt and Dave Clarke for help with the mass specs; Dr. David White for his help with bacterial culture and use of his facilities; the Barlow and Campopiano research groups for letting me use various equipment in their labs; Dr. Andrew Cronshaw for his help with MALDI at the very start of my PhD; Dr. Martin Barrios for help with the HPLC; Nicos Angelopoulos

for help with computer programing and XCMS, Dr. Thierry Le Bihan for useful discussions and his help with mass spectrometric analysis; Dr. Juri Rappsilber for letting me use his Mascot search engine and for very useful constructive criticism and advice on biomarkers. I would also like to thank two undergraduate students who worked with me: Adam Lewington (summer student) and Laura Kitchen (an honours master student) who helped me with sputum proteomics.

In addition to the Edinburgh University support, I received help from many individuals in other institutions whom I would like to thank for their interest, encouragement and constant support. Many gave me useful practical help: Drs. Michael Cannon and Gordon Davidson at P&G obtained saliva samples and gave me commercial insight into the area of 'oral care'; Drs. David Spratt, Lena Ciric and Professor Steven Porter at Eastman Dental Institute arranged for saliva samples to be collected during an induced gingivitis clinical trial and invited me to take part in useful discussions; Robert Gray and Mags Imrie at the Western General Hospital on Edinburgh gave me permission to use their lab and introduced me to SELDI; Dr. Marco Ruijken helped me with MS-Xelerator and let me share in useful discussions; Jonathon Street and Dr. James Dear at Edinburgh Royal Infirmary; Drs. Martin Wells and Martin O'Gorman at Non-linear dynamics for help with Progenesis.

I am most grateful to Dr. Perdita Barran for arranging for me to spend a week in the lab of Professor Joe Loo to learn techniques from members of his research lab at University of California, Los Angeles. They were most welcoming and generously shared their expertise, teaching me a great deal at the very start of my PhD. In particular I would like to thank Drs. Pinmanee Bootheung and Prasanna Ramachandran, and also Professor David Wong for useful discussions. I value very much the input I have had from Professor Loo and the kindness he has always shown me.

I have already mentioned the kindness shown towards me by Professor Steven Gygi for accepting me into his group as a Kennedy scholar in the first instance, and then for allowing me to continue the research towards my PhD project by arranging for my time in the US as a PhD student to be extended a further year. It has been a privilege to work alongside members of his group and I am grateful to all of them for making me feel so welcome, for sharing their knowledge and expertise always so generously. Their enthusiasm has been a spur, their friendship an invaluable bonus. I thank them all, named and unnamed, who helped me in any way.

I am indebted in particular to Dr. Kazuishi Kubota for teaching me a great deal by drawing on his own wide experience in the field of kinases. I could not have wished for a kinder, wiser and more patient mentor in those first months of my time in the Gygi lab. Through him I had the best possible introduction to kinases and computer programming I could have had because he was always ready with assistance whenever I needed it.

Dr. Willi Haas was equally generous in the time he spent teaching and training me in many skills and techniques new to me at the outset; even more importantly, he continued teaching me by answering my many questions with a great deal of patience and goodwill throughout the whole period. Without his help it would have taken me much longer to assimilate all the new information. Dr. Mat Sowa was similarly generous in the time he spent introducing me to the world of molecular biology and helping me with cloning. In the KAYAK team I have all its members to thank for their help, their spontaneous wit, their friendship and the guidance they have always been happy to give me. I would especially like to thank Dr. Ryan Kunz for his ready assistance whenever I needed his help, advice and expertise. Expert help and ready kindness came too from Drs. Noah Dephoure, Woong Kim, Ed Huttlin and Judit Villen. I would also like to thank Ramin Rad, Deepak Kolippakkam and Julian Mintseris for computational help.

Dr. Lily Ting has been much more than a fellow team member and good colleague; her friendship has been hugely appreciated in the past year. Not only has she given me much practical help by reading over chapters, she has given me much useful and thoughtful advice drawn from her own experience.

I thank friends and relatives back in the UK whose support has been much appreciated over the years. Sanaul Siddiquee was an extremely kind and very supportive friend. From his experience of life and work he offered wisdom and friendship of an invaluable kind. To Myriam Scansetti I am also grateful for being a very good friend and fun flatmate while we both pursued our PhD projects at Edinburgh.

My aunt, Jean McAllister, used her experience as a clinical biochemist to advise me from quite early on about different career paths in chemistry, and enabled me to gain insight into the nature of her line of work. She always had time to answer my many questions and provided me with much sensible and practical advice. With a quite different professional and business background my uncle, Geoffrey Wainwright, has taught me a great deal about personal effort and integrity in all aspects of work, and I have learned a great deal about industry and enterprise through conversations and discussions over the past five years in particular.

I thank Alex for all his patient help and kindness particularly as my thesis approached completion and deadlines became ever pressing; his eager and generous spirit always wills me on. I feel very grateful for the the unseen kindness and constant support of my parents, Catherine and Hugh and of my brother Iain and sister-in-law, Mariola. They have no idea how much I appreciate it.

The Caledonian Research Foundation scholarship, through The Carnegie Trust of Scotland, has funded this PhD. Additional industrial sponsorship of lab costs was generously provided by Procter and Gamble. The Kennedy Memorial Trust scholarship funded my first year in Professor Steven Gygi's research group. I would also like to thank Professor Steven Gygi's research group for additional support. To all these various funding bodies which helped finance my post graduate studies I am immensely grateful.

To CHIMA

(Catherine, Hugh, Iain, Mariola and Alex)

Table of contents

List of Figures

List of Tables

Abbreviations

Acronym	Definition
µL	Microlitre; 1 µL = 1 x 10^{-6} litres
2D PAGE	Two-dimensional polyacrylamide gel electrophoresis
2DGE	Two-dimensional Gel electrophoresis
3D	Three-dimensional
Ab	Antibody
ACN	Acetonitrile
AEX	Anion exchange
Ambic	Ammonium bicarbonate
ANOVA	Analysis of variance
APC	anaphase promoting complex
APPs	Acute-phase proteins
AQUA	Absolute Quantification of Proteins
ATP	Adenosine triphosphate
BAL	Bronchoalveolar lavage
BKL	Bioknowledge Library
Blast	Basic Local Alignment Search Tool
BN	Blue native
BP	Biological process
BPC	Base peak chromatogram
BSA	Bovine serum albumin
BW	Bronchial washing
CapLC	Capillary liquid chromatography
CDFF	Constant depth film fermentor
CF	Cystic fibrosis
CID	Collision induced dissociation
CKM	Creatine phosphokinase
cm	Centimetre; 1 cm = 0.01 metres
CNS	Central nervous system
COFDRC	Collaborative Oral Fluid Diagnostic Research Centre
COPD	Chronic obstructive pulmonary disease
COW	Correlation optimised warping
CP	Core particle
CSF	Cerebrospinal fluid
CST	Cell Signaling Technologies
C-terminus	Carboxyl terminus
Da	Dalton; 1 Da = 1.6605 x 10^{-27} kilograms
DIGE	Differential In Gel Electrophoresis
DMEM	Dulbecco's Modified Eagle Medium
DNA	Deoxyribonucleic Acid
DTT	Dithiothreitol
e	Electron charge; e = 1.602 × 10^{-19} C
ECD	Electron capture dissociation
EDI	Eastman Dental Institute
EDTA	Ethylene diamine tetra-acetic acid
EGF	Epidermal growth factor
EGTA	Ethylene glycol tetraacetic acid
EIA	Enzyme immunoassay
EIC	Extracted ion chromatogram
ELISA	Enzyme-Linked Immunosorbent Assay
ERI	Edinburgh Royal Infirmary
ESI	Electrospray ionisation
ETD	Electron transfer dissociation
EtOH	Ethanol
FA	Formic acid
FBS	Fetal bovine serum
FDR	False discovery rate
FPLC	Fast protein liquid chromatography
FTICR	Fourier transform ion cyclotron resonance

FT-LTQ	Fourier transform linear trap quadrupole
GAG	Glycosaminoglycan
GCF	Gingival crevicular fluid
GeLC-MSMS	Gel electrophoresis liquid chromatography tandem mass spectrometry
GFP	Glu-fibrinopeptide
GO	Gene Ontology
HAI	Human Antibody Initiative
HBPP	Human Brain Proteome Project
HEK	Human embryonic kidney
HGPI	Human Disease Glycomics/Proteome Initiative
HIV	Human immunodeficiency virus
HLPP	Human Liver Proteome Project
HMP	Human Microbiome Project
HMS	Harvard Medical School
HOMD	Human Oral Microbiome Database
HPLC	High performance liquid chromatography
HCT	High capacity trap
HUPO	Human Proteome Organisation
i.d.	Inner diameter
IA	Iodoacetamide
ICAT	Isotope-coded affinity tagging
IDA	Iminodiacetic acid
IEF	Isoelectric focusing
IEX	Ion exchange chromatography
IgA	Immunoglobulin A
IL	Interleukin
IMAC	Immobilised metal affinity chromatography
IPI	International Protein Index
ITRAQ	Multiplexed Isobaric Tagging Technology
K	Lysine
KAYAK	Kinase Activity Assay for Kinome Profiling
kDa	Kilodalton; 1 kDa = 1000 Daltons
KEGG	Kyoto encyclopedia of genes and genomes
kV	Kilovolt; 1 kV = 1000 volts
L	Litre
LB	Lysogeny broth
LC	Liquid chromatography
LC/LC-MS/MS	2-dimensional liquid chromatography tandem mass spectrometry
LCMS	Liquid chromatography mass spectrometry
LC-MS/MS	Liquid chromatography tandem mass spectrometry
LEMS	Lambert-Eaton Myasthenic syndrome
LPA	Lysophosphatidic acid
LTQ	Linear trap quadrupole
M	Molar
m/z	Mass-to-charge ratio
MALDI	Matrix-assisted laser desorption ionisation
MeOH	Methanol
MF	Molecular function
MGF	Mascot Generic Format
mL	Millilitre; 1 mL = 0.001 litres
MMHD	Mouse Models of Human Disease
MMP	Matrix metalloproteinase
MOAC	Metal oxide affinity chromatography
Mr	Relative molecular weight
mRNA	Messenger RNA
ms	Millisecond; 1 ms = 0.001 seconds
MS	Mass spectrometry / mass spectrometer
MS/MS	Tandem mass spectrometry
MSX	MS-Xelerator
MudPIT	Multidimensional protein identification
nESI	Nano-electrospray ionisation

NIDCR	National Institute of Dental and Craniofacial Research
NIH	National Institute of Health
NLF	Nasal lavage fluid
NMR	Nuclear magnetic resonance
NSAF	Normalized spectral abundance factor
N-terminus	Amine terminus
o.d.	Outer diameter
OFMNC	Oral Fluid MEMS/NEMS chip
OFNASET	Oral Fluidic Nano Sensor Test
OMIM	Online Mendelian Inheritance in Man
OPLS	Orthogonal-PLS
ORFs	Open reading frames
OSCC	Oral squamous cell carcinoma
P	Parotid
P&G	Procter & Gamble
PADB	Proteomic Analysis DataBase
PAGE	Polyacrylamide gel electrophoresis
PBD	Polo-box domain
PBS	Phosphate buffer saline
PCA	Principal component analysis
PCR	Polymerase chain reaction
pI	Isoelectric point
PLS	Partial least squares
PLS-DA	PLS-discriminant analysis
PMF	Peptide mass fingerprinting
PMN	Polymorphonuclear
PPID	Protein-Protein Interaction Database
PPP	Plasma Proteome Project
PRPs	Proline-rich proteins
PSI	Proteome Standards Initiative
pSS	Primary Sjogren's syndrome
PTM	Post translational modification
QTOF	Quadrupole time-of-flight
R	Arginine
RF	Radio frequency
RIA	Radioimmunoassay
RID	Radial immunodiffusion
RIPA	Radioimmunoprecipitation assay
RNA	Ribonucleic acid
RP	Reversed phase, regulatory particle
RPTC	Rusham Park Technical Centre
RPW	Reference peak warping
S/N	Signal-to-noise
SC	Spectral count
SCIBS	Selective Chemical Intervention in Biological Systems Initiative
Scripps	University of Southern California
SCX	Strong cation exchange chromatography
SD	Standard deviation
SDS	Sodium dodecyl sulfate
SDS-PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
SELDI MS	Surface-enhanced laser desorption ionisation mass spectrometry
SiCK	Scambled siRNA
SILAC	Stable isotope labelling with amino acids in cell culture
SIMCA	Soft independent modelling of class analogy
siRNA	Small interfering ribonucleic acid
SL	Sublingual
SM	Submandibular
SN	Supernatant
SPKB	Salivary Proteome Knowledge Base
SS	Sjogren's syndrome
STK	Serine/threonine kinase

T	After toothbrushing
TBST	Tris buffered saline with tween
TFA	Trifluoroacetic acid
TIC	Total ion count, Total ion chromatogram
TK	Tyrosine kinase
TLC	Thin layer chromatography
TOF	Time-of-flight
tRNA	Transfer RNA
UCL	University College London
UCLA	University of California Los Angeles
USC	University of Southern California
UV	Ultraviolet
V	Volts
WGH	Western General Hospital
WU	Wake-up

Publications

Papers

Chaperone-mediated pathway of proteasome regulatory particle assembly. Roelofs, J., Park, S., Haas, W., Tian, G., McAllister, F. E., Huo, Y., Lee, B., Zhang, F., Shi., Y., Gygi, S. P., Finley., D. *Nature* 459, 861-865 (2009). (Appendix A.1 and section 6.2.2)

Prokayrotic ubiquitin-like protein (Pup) proteome of Mycobacterium tuberculosis. Festa, R., McAllister, F. E., Pearce, M. J., Minstseris., Burns, K. E., Gygi, S. P. *PLoS One*, 5, e8589 (2010). (Appendix A.2 and section 6.2.2)

Selective Chemical Intervention in the Proteome of *Caenorhabditis elegans*. Husi, H., McAllister, F. E., Angelopoulos, N., Butler, V. J., Bailey, K. R., Malone, K., MacKay, L., Taylor, P., Page, A. P., Turner, N. J., Barran, P. E., Walkinshaw, M. *Journal of Proteome Research.* Manuscript accepted (2010). (Appendix A.3 and section 6.2.2)

Supramolecular chemistry in metal recovery; H-bond buttressing to tune extractant strength. Forgan, R. S., Wood, P. A., Campbell, J., Henderson, D. K., McAllister, F. E., Parsons, S., Pidcock, E., Swart, R. M., Tasker, P. A. Chemical Communications, 46, 4940-4942 (2007).

Collision induced dissociation (CID) to probe the outer sphere coordination chemistry of bis-salicylaldoximate complexes. Roach, B. D., Forgan, R. S., Tasker, P. A., Swart, R. M., Campbell, J., McAllister, F. E., Duncombe, B. J., *Dalton Transactions*, 39, 5614-5616 (2010).

Conference abstracts

58th American Society for Mass Spectrometry Conference, Salt Lake City, USA, May 2010.
Poster presentation: KAYAK: Kinase ActivitY AssaY for Kinome analysis.

BioFusion, Procter&Gamble internal conference, Cincinnati, USA, October 2009.
Poster presentation: Saliva and KAYAK: Kinase ActivitY AssaY for Kinome analysis.

8th European Saliva Symposium, Egmond aan Zee, The Netherlands, May 2008.
Poster presentation: The Salivary proteome: Developing Tools for Disease Diagnostics.

European Biomarkers Summit and Proteomics Europe, Amsterdam, The Netherlands, September 2007.
Poster presentation: The Salivary proteome: Developing Tools for Disease Diagnostics.

Bristish Mass Spectrometry Society, Edinburgh, UK, September 2007
Poster presentation: The Salivary proteome: Developing Tools for Disease Diagnostics.

1 Introduction

1.1 Project overview and aims

Until very recently medical practice has been based on generalised therapy in which a standard set of treatments are used to address a particular disease[1-2]. Only recently has the idea of 'personalised medicine' become feasible in which therapies are specifically tailored to an individual's genetic makeup to yield optimal results[3]. With the dramatic reduction in the cost of DNA sequencing, in the future medical treatment in the developed world will become increasingly personalised and effective[4-5]. Whilst the use of genetic markers are useful for gauging a patient's potential responsiveness to various drugs or understanding genetic disorders, knowledge of an individual's gene sequence alone is often inadequate for understanding and treatment of disorders related to transcriptional regulation, protein expression levels, and abnormal enzymatic activity[6-7]. Protein biomarkers are able to detect abnormal cellular dynamics not revealed by DNA sequencing, and are likely to be a critical tool in personalised medicine[8]. Protein profiling and biomarkers are likely to play an increasingly significant role in screening, diagnosis, prognosis, and monitoring of patients in a wide variety of diseases. Classical proteomics focuses on protein identification and abundance whilst protein activity has been relatively unexplored[9]; the assessment of both will be necessary for fully understanding the human proteome. Drug development and testing in pharmaceutical companies may also be significantly benefited by continued progress in these fields[10-11].

It is important that any tool used for screening, diagnosis, or prognosis should use easily obtainable samples that provide robust information. Saliva, as an easily accessible body fluid, fits these criteria and holds great promise as a diagnostic tool[12]. Despite the ease of access and non-invasive nature of saliva collection, the majority of biomarker efforts have been directed towards blood, and the salivary biomarker field has been relatively unexplored. Furthermore, most protein biomarkers have been validated, to date, using protein abundance approaches[8, 13] and there is a great need for technologies that can comprehensively assess protein activities that may be more indicative of disease state than protein abundance[9].

This project analysed saliva using both proteomics and protein activity profiling in order to discover disease biomarkers. In this thesis, the salivary proteome was characterised and biomarkers were investigated from an induced gingivitis clinical trial. Developments and novel applications of KAYAK (Kinase ActivitY Assay for Kinome profiling) were investigated. This assay quantifies one specific type of protein activity (kinases) and offers great potential for cancer-diagnostics, treatment options, and lead compound drug evaluation. The strategy used in this kinase assay can also be applied to many enzyme classes and provides a promising new technique for activity-based biomarker discovery.

The specific project aims were:

i) To investigate the salivary proteome comprising both the supernatant and residue portion, including proteins of human and bacterial origin (Chapter 3).

ii) To develop and apply a mass spectrometry-based multiplexed kinase assay and kinase activity/kinase abundance correlation profiling to a variety of different systems including saliva (Chapter 4).

iii) To identify protein-abundance and protein-activity salivary biomarkers of gingivitis (Chapter 5).

Prior to addressing these aims, a general overview of proteomics, mass spectrometry, saliva, and biomarkers will first be discussed.

1.2 Proteomics

The proteome can be defined as the total protein complement of an individual cell, tissue or organism under a given set of conditions[14]. Proteomics involves any large scale or systematic characterisation of the proteome, and represents the functional genome[15]. The aim of proteomics is to generate an overview of the characteristics and functions of proteins in a given biological system, at any particular time, to reveal information on cellular dynamics and cell-to-cell crosstalk[16-17]. The proteome is dynamic – proteins are modified and change in abundance and sub-cellular location with time. The dynamism of proteins is dependent on the state of the cell (metabolic, activation, pathological) and internal and external signals. The proteome reflects the state of the cell at the time of sampling, thus revealing the molecular mechanisms behind the control of all aspects of cell biology.

Proteomics can be divided into two types in terms of scale: global and targeted proteomics[18]. The goal of global proteomics is to analyse all the proteins present in a cell. An extension of global protein analysis can involve examining differences in protein profiles under different conditions. Since differences will always be observed, the main challenge is to determine those that are biologically significant. On the other hand, targeted proteomics attempts to characterise a well-defined subproteome, for example phosphoproteins, glycoproteins, cell surface proteins, individual organelles, macromolecular complexes, and cellular machines[19-20]. The main challenge is the difficulty in reproducible isolation of subproteomes and differentiating relevant proteins from contaminants[20-21].

1.2.1 Protein identification by mass spectrometry

There are two main approaches for protein identification using mass spectrometery (MS): ‘top-down’ and ‘bottom-up’. In the top down approach intact proteins are analysed directly by MS, whilst in the

bottom-up method the proteins are first digested into peptides before MS analysis. The main advantages of the bottom-up approach for sequencing are that the currently existing database search engines are more developed for the analysis of bottom up fragments, determination of the mass of peptides is more accurate than measuring the mass of whole proteins, and the bottom-up approach allows for higher throughput[22]. Only bottom-up approaches will be considered here. There are two main routes for protein identification using bottom-up approaches: Peptide Mass Fingerprinting (PMF) and tandem MS.

1.2.1.1 Protein identification using peptide mass fingerprinting

In the early days of proteomics (mid 1990s), peptide mass fingerprinting was used to identify proteins by obtaining and matching the masses of a group of peptides derived from a protein by sequence-specific proteolysis against the masses of the peptides from *in silico* digested proteins in a database[23-24]. Protein digestion by a sequence-specific protease generates peptides with predictable masses, which are generally unique for a specific protein, thus providing a mass fingerprint for the protein. Trypsin is often the protease of choice, and it cleaves proteins at the C-termini of arginine (R) and lysine (K) residues. MS analysis of peptides results in a list of mass/charge (m/z) ratios. The experimentally derived m/z values are compared to a database of *in silico* digested proteins, generated using the same sequence-specific protease. An algorithm is then used to compare experimentally derived m/z ratios of peptides versus theoretically predicted peptide m/z ratios for each protein in database. Scores are assigned depending on the quality of the match and the number of matches of peptides to the protein. PMF is generally performed on MALDI TOF (Matrix Assisted Laser Desorption Ionisation Time Of Flight) mass spectrometers. In MALDI the analyte is embedded in a matrix that is dried and then volatilised in a vacuum under UV laser irradiation. It is often combined with a TOF analyser, where the mass-to-charge ratio of an analyte ion is deduced from its flight time through a tube of specified length under vacuum.

1.2.1.2 Protein identification using tandem MS

The use of peptide fragmentation spectra from tandem mass spectrometers has now replaced PMF as the main protein identification approach. Tandem mass spectrometers have two (or more) stages of mass analysis and they are able to fragment peptide ions and record the resulting peptide fragmentation spectra. There are various ways by which peptides can be fragmented, depending on the type of mass spectrometer, and this in turn dictates at what point in the peptide backbone the peptides are fragmented. The most common fragmentation method is CID (Collision Induced Dissociation) and is the usual process for triple quadrupole, ion trap, quadrupole-time of flight (QTOF) hybrids and LTQ Orbitrap mass spectrometers (discussed in more detail in Appendix B). In general, the peptide ion of interest is isolated from the rest of the ions from the source. The isolated peptide ion is then fragmented in a collision cell by collision with a non-reactive gas, such as helium or argon, and the fragment ions detected. The process of CID on each of the peptide ions can be automated. For complex mixtures of proteins, it is necessary to have very low flow rates (50-300 nL/min) into the mass spectrometer to allow time to isolate each of the peptides and to perform CID on them. Protocols

for automated, instrument-controlled precursor ion selection have been developed in which, in a particular time-frame, the ion with the greatest intensity is selected (data-dependent acquisition MS/MS) and CID is performed on it.[25] The next most intense ion is then selected and so on.

The majority of fragment ions from low-energy CID result from cleavage of the amide bond and are referred to as b and y ions (Figure 1.1). A b ion is formed when the positive charge associated with the parent peptide ions is retained by the amino terminal fragment side, whilst a y ion is formed when the charge is retained on the carboxy terminal fragment of the fragmented amide bond[26]. The subscript for b and y ions represent the number of amino acid residues present on the fragment ion counted from the amino and carboxy termini respectively[27].

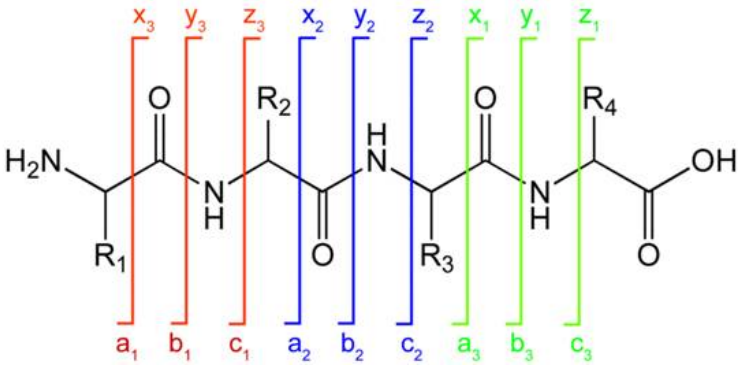


Figure 1.1 Chemical structure of a peptide with the designation for fragment ions (Roepstorff-Fohlmann-Biemann nomenclature)[26-27] that can be generated following gas phase fragmentation. Peptide fragmentation through collisions with residual gas generally leads to the generation of b and y ions. Schematic adapted from Faull *et al.*[28]

The information in the peptide fragmentation spectra can be used to identify proteins in two different ways. Firstly, *de novo* sequencing: amino acid residues can be determined from the mass difference of successive fragment ions of the same type (e.g. $b_n - b_{n-1}$)[29]. The advantage of *de novo* sequencing is that the search is not restricted to peptides contained in a database. However, there are several problems with this approach as it is generally difficult to correctly identify the whole amino acid sequence and the results are time-consuming to analyse manually. Another difficulty in *de novo* interpretation is that when using tandem MS spectra, not all fragment ions are detected and many of the intervening peaks may or may not belong to the series. Although there have been some attempts to automate this process, it remains an extremely difficult task[30,31]. The second approach for protein identification is the comparison of the m/z peaks in the CID spectra with theoretical *in silico* generated MS/MS CID spectra. This approach is easily automated and is the more prevalent method used today. Only a small fraction of the total number of possible peptide amino acid sequences occur in nature and therefore, whilst a tandem mass spectrum may not contain enough information for *de novo* sequencing, it often contains sufficient information to be matched to a peptide sequence in the database based on the observed m/z values of the fragment ions.

1.2.2 Quantitative proteomics

To understand biological systems and their perturbations, measurement of changes in protein abundance as well as protein identification is important[32]. Quantitative proteomics involves the global analysis of changes in protein abundance in a particular system (cell, tissue or organism) under different conditions/treatments. Protein abundance can be determined using either relative (to other proteins in an experiment) or absolute quantitation, but in general, relative quantitation is more commonly used in large-scale proteomics experiments on account of ease and cost. In gel-free proteomics experiments, quantitation is normally calculated at the peptide level, rather than the protein level. However, the signal intensity of a particular peptide is not directly proportional to the protein amount due to varying ionisation efficiencies of different peptides, protease accessibility, and peptide solubility[22]. There are several approaches for quantitation using mass spectrometry (Figure 1.2)[33] including stable isotope labelling[34-35] and label-free quantitation based on either the ion intensity[36] or spectral count[37-38].

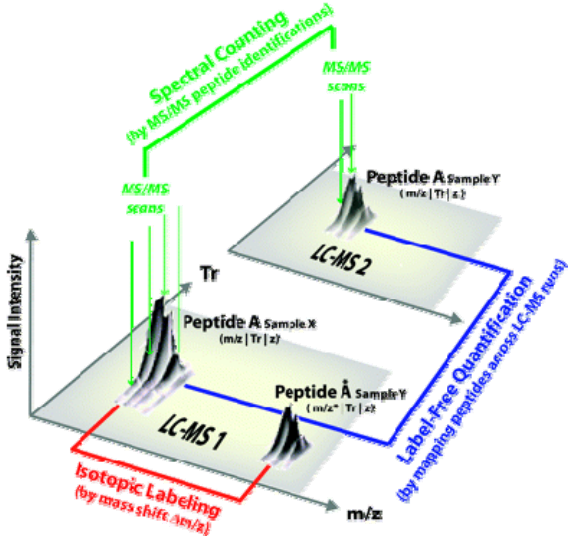


Figure 1.2 Schematic of the three main quantitation approaches in LC-MSMS proteomics. Quantitation using isotopic labelling (shown in red) uses the abundance of peptide A (sample X) versus the abundance of the heavy isotope labelled peptide A* (sample Y) for relative quantitation. Quantitation using spectral counting (shown in green) relies on the number of times a peptide was successfully sequenced as a measure of peptide/protein abundance and compares this across different LCMS runs (sample X and sample Y). Quantitation using label-free ion intensity (shown in blue) compares the ion intensity of peptide A (sample X) with the same peptide in a different sample (Sample Y) in a different LCMS run based on matching up of the m/z and the retention time. Adapted from Mueller *et al.* [33]

There are various strategies for quantification using LC-MSMS (liquid chromatography tandem mass spectrometry) (metabolic labelling, chemical modification, label-free) depending on the point in an experiment the quantitative information is generated, and on when the label is added. For example, metabolic labelling can be used to label during cell culture, chemical modification can label at either the protein or peptide level, whilst label-free methods rely on the mass spectrometer extracted ion intensity, or the spectral count, and require no labelling.

1.2.2.1 Stable isotope labelling

Quantitative proteomics using stable isotope labelling was first reported in 1999 by several research groups[39-40]. Stable isotope labelling theory proposes that the relative signal intensity in a mass spectrometer of two analytes that are chemically identical but of different stable isotope composition (and thus distinguishable in a mass analyser) are a true representation of the relative abundance of the two analytes in the sample[41]. The differential stable isotope labelling approach allows relative peptide quantitation between two (or more) samples in the same LC-MSMS run. Two samples are differentially labelled with a light and heavy isotope and can be detected by their mass difference ($\Delta m/z$), resulting in a relative quantitation of the protein amounts. Absolute quantification using stable isotopes is also possible by comparison with a labelled synthetic peptide analogue that acts as an internal standard (e.g. AQUA)[42], but will not be discussed further in this section.

ICAT (isotope coded affinity tag)[39] and iTRAQ (isobaric tag for relative and absolute quantitation)[43] are chemical modifications where the samples are treated with heavy and light forms of a reagent. Chemical modifications can be performed either on proteins or peptides, depending on the actual tag. ICAT analysis involves labelling proteins at cysteine residues with heavy and light ICAT labels prior to trypsin digestion. The ICAT reagent is attached to cysteine residues by a cleavable biotin group which allows rapid affinity isolation of cysteine containing tryptic peptides on an avidin column[44]. The heavy and light ICAT tag results in a 8 Da mass shift between the respective heavy and light peptides, and through comparing the peak area of the heavy and light peptide pairs in the MS survey scan can allow relative quantitation[44]. One of the disadvantages of ICAT is that only cysteine containing proteins can be analysed, although this can be advantageous in the analysis of a complex sample where the simplification of the analysis can often result in greater sensitivity[45].

The iTRAQ system is similar to ICAT but the labelling is performed following trypsin digestion and peptides, rather than proteins, are tagged. The amine group of a peptide is labelled with isobaric labelling reagents and fragmentation of the tag attached to the peptide generates reporter ions of masses 113-119 and 121 Da which are unique to the tag used[46]. By comparing the relative intensity of the reporter ions it is possible to obtain relative quantitation of the original protein amounts[46-47]. There are currently 8 iTRAQ labels, allowing 8-plex quantification[47]. Advantages of iTRAQ include multiple comparisons of different samples/treatments in the same experiment along with complete sample labelling[48]. A disadvantage of iTRAQ is that the chemical tag is introduced at a late stage of the experimental workflow. This means that many sample preparation steps must be performed prior to the labelling, increasing the risk of inadvertent errors due to differences in sample preparation. Another labelling technique is to use $^{16}O/^{18}O$ digestion labelling where peptides are labelled during or following trypsin digestion in the presence of $H_2{}^{16}O$ or $H_2{}^{18}O$[49].

Metabolic labelling, where the label is introduced *in vivo* during growth, can be advantageous since sample preparation errors are reduced as the samples are isotopically labelled very early in the procedure prior to protein extraction, and hence any sample preparation errors performed post

labelling affect both the heavy and light labelled samples in parallel[50]. In general, metabolic labelling using $^{14}N/^{15}N$ is most commonly used and has been performed on many organisms including bacteria[40], yeast[51], plants[52], cell culture[53], *Drosophila melanogaster*[54], *Caenorhabditis elegans* and rats[55]. The main disadvantage of metabolic labelling is that samples can only be labelled during growth, thereby excluding body fluids, clinical biopsies, and environmental samples from being analysed. Furthermore, the aim of achieving complete labelling can often result in long culture times, and incomplete labelling is common for complex multicellular organisms[55].

One popular metabolic labelling procedure is SILAC (Stable isotope labelling with amino acids in cell culture) which utilises two cultures of cells.[56] One group of cells is cultured with unlabelled (natural isotopic abundance) amino acids and the other culture is fed with amino acids in which one amino acid, generally arginine or lysine, has been replaced with an isotopically (often ^{13}C and/or ^{15}N) labelled form. As the cultures grow, the cells incorporate the natural or heavy isotopes into their newly synthesised proteins. Following proteolytic digestion the ratio of peak intensities or peak areas in the mass spectra can be used to determine the relative abundances of the proteins. A disadvantage of SILAC, compared to $^{14}N/^{15}N$ labelling, is the possibility of quantitation errors through *in vivo* interconversion of isotopically labelled amino acids. Arginine is converted into proline in some cell lines and glycine can be converted to various amino acids[57-59] leading to partial loss of labelling and unpredictable mass shifts.

The main advantage of labelling methods is that they allow the relative quantitation of peptides/proteins from different samples in the same LCMS run, meaning that the method does not suffer from multiple measurement noise or errors from misalignment of chromatographic traces. For label-free analysis it is imperative that the chromatographic and mass spectrometry conditions are identical between runs. Disadvantages of labelling based quantitation include high reagent cost, increased time and complexity of sample preparation, difference in ionisation efficiencies of labels, incomplete labelling, the potential requirement of specific quantitation software, and the limitation of comparing only two samples. Consequently for experiments requiring the comparison of large numbers of samples, such a biomarker discovery, there has been increased interest in the application of label-free proteomics for quantitation.

1.2.2.2 Label-free quantitation

Label-free quantitation encompasses methods that use spectral counting as a measure of protein abundance, as well as methods that directly compare the signal intensity between different LCMS runs. The label-free approach is significantly less costly than the labelled approach because high-cost, heavy isotopic label tags are not necessary. As a result, the label-free approach is more suited comparing a range of different treatments in clinical samples and allows for retrospective comparison.

1.2.2.2.1 Ion intensity

Label-free quantitation based on the extracted ion intensity (EIC) quantifies proteins in separate

LCMS or LC-MSMS analyses through aligning the chromatographic total ion count (TIC) traces from MS survey scans and comparing the peptide relative peak intensities (peak areas or heights)[36, 60-62]. A challenge in label-free peptide ion intensity quantitation for complex peptide mixtures is that changes in the elution of the peptides may affect the quantitation. It is therefore critical that the same data acquisition conditions are performed for all experiments: same mass spectrometer method, same LC method, same column and temperature[63]. This becomes more problematic where multi-dimensional chromatographic separations are used, since even small variations in the chromatography will lead to irreproducible peptide separation[64]. Matching of peptide peaks between different MS analyses can be challenging on account of variations in both the LC retention time and the m/z[65]. The changes in LC retention time of the peptides are often not linear and accurate correction is difficult[64]. Minimal sample handling is preferred to reduce variations between samples and minimise errors[65].

Much of the published research using label-free ion intensity quantitation uses spectra with high resolution and high mass accuracy from FT-LTQ, LTQ Orbitrap and QTOF instruments with few reports using low resolution mass spectrometers such as ion traps[63]. Accurate matching of peptides between different runs is easier with greater accuracy m/z and the software packages are often optimised for the higher resolution mass spectrometers[66]. However, label-free ion intensity quantitation software is often designed to work with a particular mass spectrometer and may only give quality results with one particular instrument setup. For hybrid mass spectrometers it is possible to acquire LC-MSMS data, and hence quantitative information from the LCMS may be combined with identification information from the MSMS scans. However, linking the LCMS with the LC-MSMS is not straightforward and there is currently much software development in this area[66]. There are a wide variety of software packages available for label-free quantitation, both open source and commercial[165,166,170,171] (Table 1.1).

Most of the software packages have fairly similar workflows. Preprocessing algorithms are employed for smoothing and despiking the data and then peptide peaks are detected through their characteristic isotopic pattern. The chromatograms are aligned to account for changes in the retention time and the TICs are normalised based on a spiked in reference or the total TIC area. The EICs of each of the peptides are then compared between the different LCMS runs. Finally, MSMS information on the peptides can be linked through the m/z and retention time (Rt) to the peptides in the LCMS run. There are two main approaches for the combination of LCMS and LC-MSMS data. Firstly, MSMS information from a single LC-MSMS run can be used when quantitation LCMS information has already been extracted. However, because in one LC-MSMS run not all the peptides have MSMS data, another approach has been to incorporate MSMS information from several LC-MSMS runs or use inclusion lists which target peptides of interest[33].

1.2.2.2.2 Spectral counting

There are various features that can be correlated with protein abundance, for example the number of identified peptides, the number of unique peptides, the sequence coverage, and the spectral count[67].

The spectral count is the sum of all the acquired MSMS spectra and represents the identified peptides for a protein[68]. In spectral counting quantitation experiments, two or more samples are analysed using LC-MSMS independently with similar data acquisition parameters. Protein identification is performed on the individual samples separately and the spectral count of each protein is then compared between samples. A variety of different normalisation procedures can be used to account for both run-to-run variations[69] and the variation in protein sizes. For example, normalisation is often performed using protein length or the expected number of proteolytic peptides. The normalised spectral abundance factor (NSAF) is a commonly used approach that effectively divides the total spectral count for a protein by its length[70-71].

One of the challenges with the spectral count quantitation method is the significant variability at low spectral counts (e.g. one to three peptides per protein)[33]. Shotgun proteomic experiments (high performance liquid chromatography combined with tandem mass spectrometry) favour the detection of the more abundant peptides, and the detection of peptides present at only low concentrations can be poor and irreproducible[72]. Also, at high spectral counts saturation behaviour is often observed[73]. Spectral counting is very dependent on the accuracy of the MSMS peptide identifications, and consequently any errors made at the peptide level propagate though to the protein level[33]. Another source of error is that one peptide can be matched to several proteins leading to problems of redundancy at the protein level.

In general, spectral counting is considered semi-quantitative, since the spectral count is an estimate of protein amount and the physicochemical properties of the peptides are not considered. Spectral counting assumes that all the peptides respond linearly with increased protein abundance, but this is not necessarily the case. Further, the peptide's size, charge and hydrophobicity have a significant effect on the efficiency of ionisation[65]. Spectral counting is able to quantify proteins from many independent experiments and is suited to samples with large dynamic ranges[38]. Low resolution instruments are also suited to spectral counting because low resolution data are not ideal for other quantitation methods. Spectral counting has been widely used in many biological samples including biomarker discovery in urine[74], saliva[75] and lung cancer[76].

Despite the lower accuracy of label-free quantitation in proteomics, the absence of a need for expensive reagents and tags/labels, a greater dynamic range, and no limitation on the number of samples for analysis, means that label-free is the preferred technique in many studies. This is particularly true for those studies involving biomarkers where the analysis of numerous samples is required[65, 70].

1.2.2.3 Software packages for proteomics quantitation

Data processing and bioinformatics are major analytical challenges in quantitative proteomics and in the last five years, there has been a rapid increase in the number of commercial and open-source software packages for quantitation of proteomics data (Table 1.1)[77].

Table 1.1 Quantitation software programs for quantitative proteomics experiments that are publicly available. Adapted from Ting *et al.*[78]

Software	MS instrumentation	Labelling strategy	Quantitation strategy	Reference
MSQuant	QSTAR, QTOF, LTQ-FT	SILAC ($^{14}N/^{15}N$ possible with extra software)	Area under centroided MS^1 XIC of isotopmers	Schulze & Mann[79]
XPRESS	LCQ, QTOF, MALDI-TOF-TOF	ICAT (also SILAC, and metabolic labelling)	Area under MS^1 XIC of isotopmers	Han *et al.*[80]
ASAPRatio	QTOF, LTQ, LTQ-FT, LTQ-Orbitrap	Stable isotopic labelling	Area under MS^1 XIC of isotopmers	Li *et al.*[81]
MFPaQ	QSTAR	ICAT, SILAC	Peak intensity of MS^1 isotopomers	Bouyssie *et al.*[82]
MaxQuant	High resolution mass spectrometers (LTQ-Orbitrap)	Stable isotopic labelling especially SILAC, label-free	Intensity of MS^1 3D-peaks (intensity *vs.* m/z *vs.* time) scans of isotopomers	Cox & Mann[83]; Graumann *et al.*[84]
RelEx	ThermoFischer (.raw)	$^{14}N/^{15}N$ metabolic labelling	Background-subtracted intensity ratio of MS^1 XIC of isotopmers	MacCoss *et al.*[51]
Census	Low and high-resolution mass spectrometers	Label-free, stable isotopic labelling, iTRAQ	MS^1 based on RelEx, or MS^2 from single-reaction monitoring (SRM) scans or iTRAQ	Park *et al.*[85]
QN	High resolution hybrid ion trap mass spectrometer; LTQ-FT	$^{14}N/^{15}N$ metabolic labelling	Area under MS^1 XIC of isotopmers	Andreev *et al.*[86]
MSight	Applied Biosystems, Brucker, Waters, ABI-SCIEX, ThermoFinnigan mass spectrometers	Label-free	2D representation of LC-MS data, similar to 2DE	Palagi *et al.*[87]
APEX	Most	Label-free	Absolute quantitation by MS^2 spectral counting, corrected by machine learning-based prior expectation of observing each peptide based on physicochemical properties	Lu *et al.*[88]
MapQuant	LTQ-FT	Label-free	MS^1 scans are converted into 2D maps (RT *vs.* m/z). Quantitation of 2D map features.	Leptos *et al.*[89]
PEPPeR	LTQ-Orbitrap, LTQ-FT	Label-free	Spectral feature quantitation by landmark and peak matching of MS^1, uses MapQuant as part of the PEPPeR pipeline	Jaffe *et al.*[90]
Masic	Most	Label-free	Accurate mass and time tag quantitation from MS^1	Monroe *et al.*[91]
msInspect/AMT	Most	Label-free	Accurate mass and time tag quantitation from MS^1	May *et al.*[92]
SpecArray	High resolution, high accuracy QTOF or similarly performing analyser	Label-free	Peptide array, quantitation of MS^1 single ion chromatograms based on ASAPRatio algorithm	Li *et al.*[93]
ZoomQuant	Ion trap mass spectrometers	^{18}O labelling	Isotopomer peak areas from MS^1	Halligan *et al.*[94]
ProQuant	QSTAR, QTrap; Applied Biosystems mass spectrometers	iTRAQ (4-plex)	Isotopomer peak area from MS^2 of reporter ions (114-117 m/z for 4-plex)	Applied Biosystems, 2004
ProteinPilot	QSTAR, QTrap; Applied Biosystems mass spectrometers	SILAC, iTRAQ (4-plex and 8-plex)	Isotopomer peak areas from MS^1 for SILAC. Isotopomer peak area from MS^2 of reporter ions (114-117 m/z for 4-plex, 113-121 m/z for 8-plex)	Applied Biosystems, 2006
Multi-Q	Applied Biosystems, Brucker, Waters and ThermoFinnigan mass spectrometers	iTRAQ	Peak intensities of reporter ions in MS^2	Lin *et al.*[95]
Libra	Most, QTOF, QSTAR, TOF-TOF	iTRAQ	Peak intensities of reporter ions in MS^2	Keller *et al.*[96]
i-Tracker	Most	iTRAQ	Area under report ion peaks in MS^2	Shadforth *et al.*[97]
Mascot	Most	SILAC, 14N/15N metabolic labelling, 18O, iTRAQ	Area under MS^1 XIC of isotopmers, area under report ion peaks in MS^2 for iTRAQ, spectral counting.	Matrix Science, 2008
XCMS	Most	Label-free	Quantitation of MS^1	Smith *et al.*[98]
MS-Xelerator	Most	Label-free	Quantitation of MS^1	MsMetrix, 2007
Progenesis LC-MS	Most	Label-free	2D representation of LC-MS data, similar to 2DE	Nonlinear Dynamics, 2008

Many of the software packages are dependent on the specific instrument type and labelling/label-free strategy. While there are a few software packages available, such as MaxQuant[83], Trans-Proteomic Pipeline[96] and Mascot v2.2 (Matrix Science, UK), in general, very few publically available software platforms exist for integrated data processing and analysis. Integrated data analysis allows consistent data processing and facilitates comparison between experiments as well as between different research groups[96].

1.2.3 Subproteomics

Subproteomics can refer to proteomics of a particular organelle, protein complex or body fluid, as well as a proteomics investigation of an affinity enriched sample such as the phosphoproteome or glycoproteome[99]. A greater depth of protein coverage may be obtained by focussing on a particular subproteome. This thesis involves the study of the salivary proteome as well as the phosphoproteome of saliva. A brief introduction to the proteomics of body fluids and phosphoproteomics is discussed below.

1.2.3.1 Proteomics of body fluids

One of the applications of proteomics is to identify disease biomarkers and, for this application, the proteomes of body fluids are likely to be the most useful and clinically relevant[100]. To achieve this aim for biomarker discovery, two phases of proteomic research have been proposed: a mapping phase and a scoring phase[100]. The mapping phase involves large scale, comprehensive analysis of the proteomics of body fluids and their subsequent organisation into databases (e.g. Global Proteome Machine[101,102] and Peptide Atlas[103,104]). The scoring phase involves quantitation of a core set of proteins that can effectively represent the whole proteome or a subproteome of particular interest[100]. Many human body fluids have been investigated using proteomics; including blood[105], which has been the most studied, urine[106], saliva[107], cerebrospinal fluid[108-112], tears[113], and seminal fluid[114]. The Human Proteome Organisation (HUPO) completed the first large scale study of human serum and plasma proteins in which 889 proteins were identified with high confidence[115] from a total of 3,020 proteins which were identified with two or more unique peptides[116]. Since blood comes into contact with all the tissues, tissue specific proteins are secreted into it and hence plasma proteomics has the potential to diagnose many diseases affecting all parts of the body[117]. However, since the volume of blood is relatively large, the abundance of proteins secreted into the blood from any particular organ can be very low and there may be greater opportunity for biomarker discovery in 'proximal fluids' which are those fluids that come into contact with only a few tissue types[100]. The concentrations of tissue specific proteins that are indicative of the health/disease state of that tissue are likely to be higher in proximal fluids and therefore these fluids offer greater diagnostic value for certain diseases[100]. For example, CSF (cerebrospinal fluid) can give an indication of the biochemical environment of the brain as it is in continuum with the extracellular fluid of the CNS (central nervous system). There are various brain-specific proteins present in low concentrations in CSF. However, more than 80% of the proteins in CSF are thought to originate from plasma[108-112]. Furthermore, obtaining blood samples, while relatively commonplace, is more difficult than obtaining samples of other body fluids such as saliva.

Urine analysis has particular relevance for the understanding of diseases that are related to the kidneys. Diseases of the kidneys and urological tract can often cause glomerular damage. For people with normal kidney function the walls of the glomerular capillaries in the kidney only allow proteins less than 65 kDa into the urine. Therefore, many renal diseases are diagnosed by the presence of larger proteins in the urine (e.g. proteinuria). Urine is a desirable fluid for analysis as it is convenient to

collect in large amounts. In a recent study, 1,543 proteins were identified from urine, with nearly half classified as membrane proteins by Gene Ontology (GO) analysis[106]. Extracellular, lysosomal and plasma membrane proteins were found to be highly represented in the urine, whereas other intracellular proteins were low. Urinary biomarker investigation has involved the study of many diseases including prostate cancer[118], renal cell carcinoma[119], bladder cancer and[120], and urothelial carcinoma[121].

The sputome (sputum proteome) is an under-studied bodily fluid on which there is little published research. This is at least partly due to the technical difficulties involved in processing such mucin-rich samples. The highly crosslinked mucin matrix and the presence of highly charged mucins makes analysis by conventional 2DGE more difficult. The most extensive study carried out to date on the sputome used a combination of 2DGE and GeLC-MSMS and identified 258 proteins, of which 191 were of human origin[122]. The proteins in sputum originate from many diverse sources such as the secretory products of the lower and upper airways, cellular products, and inflammatory cell derived products. Proteins from 'contaminants' such as saliva, epithelial cells and gastrointestinal products are also present. An analysis comparing the sputome with the proteins of BAL (bronchoalveolar lavage), NLF (nasal lavage fluid) and saliva found that the proteins in sputum overlap significantly with those found in BAL and saliva. It is not clear, though, whether the proteins identified in both sputum and saliva are due to contamination of sputum with saliva or whether it is possible that these proteins are produced independently in the lower airway mucous glands. Many more proteins have been identified in BAL than in sputum: 1,375 versus 258[122]. A possible explanation is that the process of BAL collection causes microvascular leakage leading to the presence of serum proteins in BAL, something which would not necessarily occur during collection of sputum[122]. Various proteomic studies have been performed on saliva and this is discussed in Section 1.4.3.

One of the disadvantages of proximal fluid proteomic analysis is the restriction to diagnosis of certain diseases and that collection can be difficult in particular cases(e.g. CSF)[100]. However, it is thought that many of the biomarkers found in proximal fluids can be identified in blood plasma if a targeted approach is used[100]. Therefore, the strategy of using proximal fluids for preliminary biomarker discovery for later use as plasma biomarkers has merit.

1.2.3.2 Phosphoproteomics

Phosphorylation is one of the most common reversible posttranslational modifications (PTM) of proteins[123] whereby a phosphate group is added to a protein molecule. Phosphoproteomics is the characterisation of the degree of phosphorylation of proteins. Phosphorylation can be observed most commonly on serine residues, and to a lesser extent on threonine and tyrosine residues. Protein kinases catalyse the phosphorylation of proteins whilst phosphatases dephosphorylate proteins. Genes coding these two enzyme types constitute around 2% of the human genome[124]. Phosphorylation plays a critical role in regulating many cellular processes including cell differentiation, many metabolic functions, and forms the basis of many signalling pathways. Phosphorylation also regulates enzyme activity, complex formation, subcellular localisation and the degradation of proteins. Chemically, the addition of a phosphate group generally makes a protein more hydrophilic.

Prior to large scale mass spectrometry-based proteomic studies, the analysis of protein phosphorylation was challenging and often only focused on one phosphoprotein. Traditional methods included purification by phospho-specific antibodies which were very time consuming to generate[123]. Another approach used ^{32}P labelled ATP in *in vitro* kinase reactions followed by the separation of ^{32}P labelled phosphorylated peptides by TLC (thin layer chromatography) and subsequent MS analysis[123]. However, *in vivo* studies of phosphorylation can generally yield more useful information than studying phosphorylation *in vitro*, since nonspecific phosphorylation may occur *in vitro*. 2DGE or DIGE (followed by MALDI MS or immuno- or western blot with specific phosphoantibodies) can be used to map the phosphoproteome because phosphorylation of a protein leads to a decrease in its pI resulting in a change of spatial location on the gel. However, this approach is not suited to quantitative analysis on account of variability in the amounts of proteins transferred to the membrane and the selectivity and affinity characteristics of the antibodies. Alternatively, direct staining of 2DGE can be used with Pro-Q Diamond™, a fluorescent phosphosensor dye, that is able to discriminate between phosphorylated and unphosphorylated proteins[125]. Another method is to compare gels treated with phosphatase to untreated gels in order detect changes in the proteome[126,127].

On account of the low abundance of phosphoproteins and/or the low phosphorylation stoichiometry of phosphoproteins, it is often necessary to perform phosphoprotein enrichment when studying the phosphoproteome[123, 128]. It is important to note that the phosphorylation sites on proteins might vary and, although it is often possible to identify the major phosphorylation sites, less abundant or highly transient phosphorylation sites might be missed.

1.2.3.2.1 Methods for phosphoprotein enrichment

For native purification of phosphoproteins/peptides there are four main strategies: immunoaffinity chromatography, immobilised metal affinity chromatography (IMAC), metal oxide affinity chromatography (MOAC), and strong cation exchange (SCX) chromatography[123]. These strategies are also used for other sample separations and enrichments but can be tailored towards phosphoprotein/peptide enrichment.

In immunoaffinity chromatography, antibodies that specifically recognise phosphorylated residues irrespective of the surrounding sequence can be used to purify phosphorylated proteins and peptides. Antiphosphotyrosine antibodies are more commonly used[123,129-130] than antiphosphoserines and antiphosphothreonines[131]. This is largely because antiphosphotyrosine antibodies are more specific than the other antibodies. Other phosphorylated amino acids can be analysed more easily using antibodies that specifically target a phosphorylated peptide sequence or consensus sequence motif for a kinase.

IMAC is the most extensively used strategy for phosphoprotein/peptide enrichment. Negatively charged phosphate groups are ligands for transition metal cations such as Fe^{3+} and Ga^{3+}. These transition metal cations are linked to a solid support, such as sepharose or poros by, for example, iminodiacetic acid (IDA). The main limitation with this approach is nonspecific binding of carboxyl groups (COO^- of acidic amino acids and C-terminal carboxy groups) to the metal cation, although this can be reduced by using a low pH (COOH favoured)[132], methyl esterification of the carboxyl group[133] and a relatively high percentage of organic solvent. Elution can be performed using sodium phosphate (pH 7.4) or EDTA (ethylene diamine tetra-acetic acid).

The MOAC method uses metal oxides such as titania and zirconia to chelate phosphorylated peptides. This is based on a similar principle to IMAC, except that it is not necessary to chelate the metal to a solid support. Similarly to IMAC, the nonspecific binding of carbonyl groups of peptides remains a challenge with the approach. Non-specific binding can be reduced by the addition of dihydroxybenzoic acid[134] or using phthalic acid[135]. Addition of dihydroxybenzoic acid is reported to reduce nonspecific binding by competition between non phosphorylated peptides and dihydroxybenzoic acid for binding sites on TiO_2[136].

SCX is a common method for the purification and fractionation of proteins and peptides, where positively charged proteins/peptides are electrostatically attracted to anionic beads. By tuning the pH and ionic strength of the buffers it is possible to favour the binding of particular proteins. At low pH (2.7) tryptic phosphopeptides have less positive charge than other tryptic peptides and therefore elute earlier from SCX columns. Although SCX can be used as a standalone technique for phosphopeptide enrichment, it is generally more useful when it is used as a fractionation technique in combination with a subsequent phosphopeptide enrichment step.

1.2.3.2.2 MS strategies for phosphoprotein analysis

The bottom-up strategy is generally used for MS analysis of phosphoproteins. The C-phosphate bond is usually cleaved more easily than the peptide backbone during CID and therefore, if performing normal MSMS of peptides, the phosphate group is unlikely to be observed attached to the fragment. There are alternative approaches to analyse phosphopeptides including precursor ion scanning, neutral loss monitoring, and neutral loss triggered MS^3 fragmentation.

Cleavage of the phosphate group (PO_3^-) from a peptide by CID in negative mode generates a peak at m/z 79 in the MS/MS spectra. Using precursor ion scanning it is possible to generate an inclusion list on which MS/MS is to be performed, which can then be applied in the next run[137]. Another method involves the detection of m/z 79 in negative mode triggering a polarity switchover to positive mode MS/MS[138-139]. Similarly, loss of a phosphotyrosine immonium ion in positive mode CID at 216 m/z can be used as a marker ion for phosphopeptides[137].

The loss of various species involving phosphate (such as phosphoric acid) from precursor ions can be used as a trigger for peptide fragmentation such as in the neutral loss monitoring approach. It is possible to determine the number of phosphate groups on the phosphopeptides by observing the number of neutral losses from the precursor ion. Also, it is often possible to determine the phosphorylation sites by observing unique product ions and the fragment ion series. In positive ionisation, loss of H_3PO_4 (phosphoric acid) corresponds to a neutral loss of 98 Da from phosphoserine and phosphothreonine residues (through β-elimination), whilst loss of 80 Da corresponds to a neutral loss of HPO_3 (metaphosphoric acid) from phosphotryrosine. A loss of 69 Da corresponds to dehydroalanine from phosphoserine and 83 Da corresponds to dehydroaminobutyricacid from phosphothreonine.

Fragment ions in the MSMS spectra can be further fragmented by MS^3 and the loss of neutral fragments (e.g. phosphoric acid) in MSMS can trigger MS^3 fragmentation. This leads to better fragment ion spectra and consequently better identification of phosphopeptides[138-139].

Fragmentation techniques other than CID can be used for the analysis of phosphopeptides. Electron capture dissociation (ECD) and electron transfer dissociation (ETD) are particularly suited to the analysis of phosphopeptides because the phosphorylated group is not generally cleaved off, allowing direct determination of the modification site[140-143]. ECD involves the capture of low energy electrons by multiply protonated ions. Following the neutralisation of the charge, the radical cations dissociate forming c- and z-type fragment ions. ECD is primarily used in FTICR MS. ETD is similar to ECD and generates c- and z-type ions[144]. The main difference is that ETD uses radical ions for the transfer of electrons whereas ECD uses naked electrons. ETD is particularly suited to the fragmentation of large peptides and high charge states.

Whilst large scale phosphoproteomic studies provide information about which sites on a protein are phosphorylated, no information regarding the identity or the activity of the kinase responsible for the phosphorylation event is obtained. This issue is addressed in the next section where a mass spectrometry based kinase activity assay is discussed.

1.3 Mass spectrometry based kinase activity assay

Classical proteomics experiments generally involve the identification of large protein sets with or without relative quantitation. The results of such proteomics experiments give an overview of the steady state chemical readout of protein abundance but there is no information regarding the activity of the proteins. For enzymes, quantifying the enzymatic activity is often of more interest than the enzyme's abundance. Whilst there has been some work involving activity-based proteomics, the majority of studies on activity-based proteomics have focussed on the use of chemical tags that selectively label the active forms of enzymes[145-149]. There is currently no large scale mass-spectrometry-based enzyme activity assay strategy established and adopted by the proteomics community. The development of KAYAK (Kinase ActivitY Assay for Kinome profiling)[150-151] for kinase activity measurements, and its derivatives for other enzyme classes, offers such a strategy for global activity-based proteomic profiling.

1.3.1 Kinases

There are more than 500 genes encoding protein kinases in the human kinome and it is estimated that at least 30% of the human proteome is phosphorylated by kinases[152-153]. Protein phosphorylation (by kinases) and dephosphorylation (by phosphatases) are among the most significant regulators of signalling pathways in cells[154]. Hyperactivation of certain signalling pathways is a hallmark of most cancers[155]. The hyperactivation can be caused by the overexpression of growth factor receptors, such as Ras and PI3K, through the structural alteration of kinases such as Src and BCR-Abl, or through the loss of negative feedback control proteins such as PTEN[156-159]. During oncogenesis, rewiring of the cell signalling network occurs and the cell may reach a state whereby it becomes self-sufficient in growth signals and no longer responds to apoptotic signals[155]. Kinases are also the targets of many drug discovery programs including diabetes, autoimmune and neurological disorders[160].

Kinases generally comprise two domains: a catalytic and a regulatory domain. The catalytic domain binds ATP and the substrate, whilst the regulatory domain interacts with other proteins that modulate the activity of the catalytic domain. The kinase catalytic domain is highly conserved because all kinases bind ATP at common sites[161]. Many kinases have two major regions in which the ATP and substrate binding sites are situated. There is in a cleft linking the two regions where an activation loop keeps the substrate close to the ATP binding site. Changes in the location and conformation of the kinase activation loop affect the activity of the kinase.

Kinase inhibitor drugs generally target the ATP recognition site or act through allosteric regulation. It is very difficult to achieve high-specificity ATP recognition site competitive inhibitors, and therefore unexpected off-target inhibition is a significant problem for pharmaceutical companies. Targeting allosteric sites is advantageous on account of higher specificity, minimising off-target effects, as well as the potential to only modulate excessive kinase activities without affecting the basal activity level[162].

Kinases are divided into tyrosine kinases (TKs) and serine-/threonine kinases (STKs), based on the substrates they phosphorylate. Kinases are further divided into eight major groups based primarily on their catalytic domain sequence (Figure 1.3)[152].

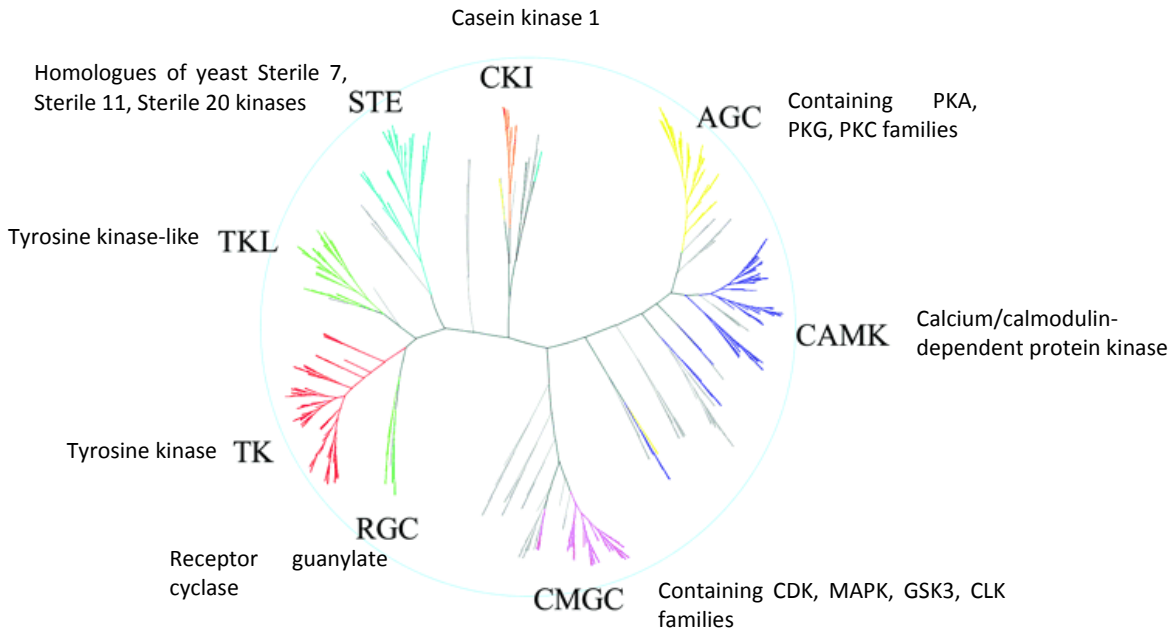


Figure 1.3 Dendogram of protein kinases illustrating the major groups. Adapted from Manning *et al.*[152]

Various strategies have been traditionally used to both infer and directly measure kinase activities[163-166]. The most common approach has been to use phosphospecific antibodies to observe phosphorylation of kinase substrates at a particular residue[167]. However, this approach does not directly measure the kinase activity. Instead, it measures phosphorylation at a specific site which is used as a surrogate of kinase activity. There is great need for high throughput techniques to screen for actual kinase activities in order to ultimately identify kinase inhibitors for drug discovery projects.

1.3.2 KAYAK: Kinase ActivitY Analysis for Kinome profiling

KAYAK is a technique recently developed by the Gygi laboratory to measure multiple kinase activities through absolute quantification of peptide substrate phosphorylation. The first version of KAYAK, reported in 2009[168], was a '90 pot' kinase reaction in which 90 synthetic peptide substrates were incubated individually with the sample (e.g. cell lysate, recombinant kinase) and ATP. Following quenching of the reaction, a known concentration of internal standard (the phosphorylated form of the substrate sequence labelled with stabled isotopes) was added. Each of the individual KAYAK reactions were then combined together, the phosphorylated peptides enriched using IMAC and then analysed using high-resolution LCMS. The absolute amount of phosphorylation is determined by comparing the peptide EIC of the light (phosphorylated substrate) peptide to the heavy (internal standard) peptide. The second modified version of KAYAK was reported in October 2009[150], and further developed the assay into a multiplexed 'one-pot' kinase reaction where all 90 peptide substrates were incubated simultaneously with the sample (Figure 1.4). There are three main advantages with the improved KAYAK assay. Firstly, significantly less lysate (90-fold) is required for the reaction. Secondly, each peptide is at 20–fold lower concentration (5 µM instead of 100 µM), which enhances the sensitivity of the assay and utilises less substrate. Thirdly, the efficiency and ease of performing many reactions simultaneously is significantly increased, which was not possible with the '90 pot' KAYAK approach.

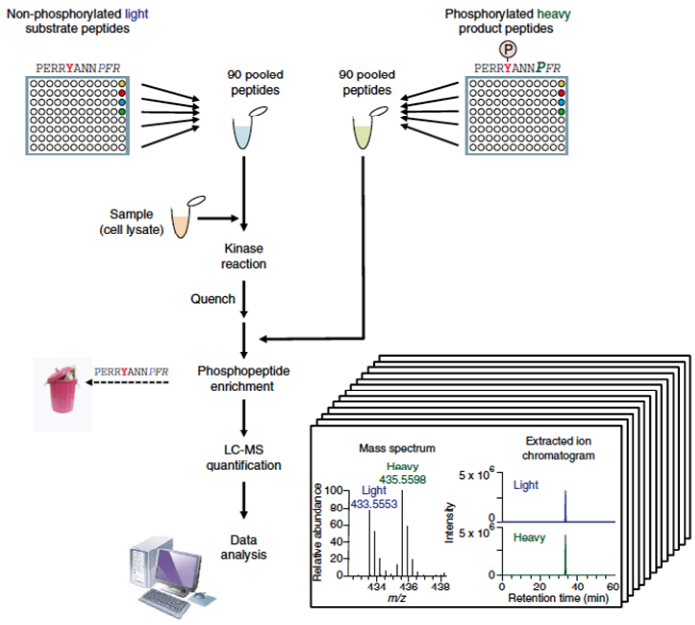


Figure 1.4 KAYAK reaction workflow using 'single pot' K90 peptide substrates. Substrate peptides are incubated with sample (e.g. cell lysate) and buffer containing ATP. The reaction is allowed to proceed for 15 min and then quenched. Internal standard peptides (stable isotope labelled phosphopeptides of the same sequence as the substrate peptides) are added at 5 µM. The phosphopeptides are enriched using immobilised metal-ion affinity chromatography and then analysed using LCMS. Figure adapted from Kubota *et al*.[150]

1.3.3 Applications of KAYAK

KAYAK can be used to determine the kinase activities in various sample types including cell lysates, tissue biopsies and recombinant kinases. By simultaneously monitoring the kinase activities of multiple kinases, an overall picture of kinase signalling pathways can be obtained.

Kinase signatures of cell lines as well as their response to different treatments can be generated. For example, KAYAK was performed on various cell lines which were treated with and without epidermal growth factor (EGF). From the upregulated kinase activities it could be seen that EGF activates the Akt/PI3K and Ras/MAPK pathways in HeLa and HEK-293 cells. As well as investigating the effect of external stimulants, KAYAK can be used to investigate the effect of kinase inhibitors. Following the success of rapamycin, a kinase inhibitor of mTOR, there has been a great increase in research into kinase inhibitors as drugs for the treatment of cancer[151]. *In vivo* there may be many off-target effects of kinase inhibitors, and the KAYAK assay offers a way to investigate which kinases and pathways may be affected by such drugs[150]. KAYAK has been performed on normal and tumour tissue from patients with renal cell carcinoma, and it was shown that there is increased activity of Akt and RSK/ERK in the tumour tissue[150]. In the future, the KAYAK profile of a biopsy could be used to determine which kinase pathways are upregulated in a patient's cancer, which can aid in determining optimum drug cocktails and provide increasingly personalised medicine.

1.3.4 KAYAK correlation profiling

Peptides can often be phosphorylated by multiple kinases. KAYAK correlation profiling is a new strategy to determine the kinase(s) responsible for a particular phosphorylation event. The sample (e.g. cell lysate) is fractionated at the protein level. Each fraction is analysed by KAYAK to determine the kinase activity, and by shotgun proteomics to determine the kinases involved. Correlation analysis of the kinase activities and kinase abundance in each of the fractions identifies kinase-substrate pairs. Shotgun analysis of the fractions identifies all proteins present and this technique may even identify regulatory partners in kinase complexes[150].

1.3.5 Comparison of KAYAK with current kinase activity assays

A range of kinase assays have been developed[169-173]. It is difficult to compare KAYAK with other techniques since there is currently not a single 'gold standard' kinase assay with which to compare it. The main competitor is possibly the method in which substrate peptides or proteins are immobilised on microarrays and the kinase activity determined with phosphospecific antibodies or radioactive ATP[173-175]. Whilst these arrays allow high throughput analysis, it is not possible to gain site specific phosphorylation information. Assays based on radioactivity have been widely adopted as they do not require the generation of expensive phosphospecific antibodies. Disadvantages of assays based on

radioactivity include a fairly high background measurement, thus hindering accurate quantitation, and the high cost of waste disposal.

In contrast, there are currently mass spectrometry methods following phosphoprotein enrichment that enable the identification of tens of thousands of phosphorylation sites. The amount of phosphorylation on these sites can even be compared between two samples for thousands of sites[176-177]. However, this analytical workflow is currently not high-throughput and requires significant amounts of starting material (in the region of 10-20mg), whilst only 20 μg per reaction is required for KAYAK.

A disadvantage of KAYAK is that since it is an *in vitro* assay, the kinase activity obtained may be misleading due to the exclusion effect of regulatory mechanisms involving protein scaffold attachment and substrate competition *in vivo*. Whilst substrate competition does exist with the 'one-pot' reaction with multiple peptide substrates, this can never mimic the *in vivo* conditions. Also, kinase function is very dependent on its localisation in a cell, where the same kinase can have very different roles in the cytoplasm and the nucleus[178-179]. In the future it may be possible to perform KAYAK *in vivo* if KAYAK peptides can be engineered to cross the cell membrane. One of the advantages of the KAYAK method is that it can be adapted to other enzyme classes, such as phosphatases and proteases and, in theory, any enzymatic reaction.

1.4 Saliva

Proteomics of body fluids is important for biomarker discovery and saliva has great potential on account of its ease of sampling[12]. A brief summary of the main components of saliva, different ways it can be collected, along with the main salivary proteomic studies that have been performed to date are discussed here.

1.4.1 Saliva composition

Saliva is an exocrine secretion that is slightly acidic (pH 6-7). It comprises a dilute aqueous solution of electrolytes, minerals, buffers and proteins and plays many physiological roles. It lubricates and cleans the oral cavity, facilitates speech, aids taste, chewing, and swallowing and aids digestion. As well as playing a significant role in oral homeostasis, it has antimicrobial activity and protects teeth by neutralising and buffering acids. It contributes to enamel pellicle formation and maintains a supersaturated solution of calcium phosphate (Figure 1.5)[180]. Some of the roles of proteins in saliva have been well studied whilst others are not known. For example, mucin is known to play several roles, including coating much of the surface in the mouth to aid lubrication, mastication and swallowing, and to protect against irritants and bacteria by regulating the adhesion of microorganisms to the oral surface[181]. Lactoferrin binds ferric iron and hence reduces the growth of certain bacteria, which require ferric iron such as carriogenic (dental caries forming) *Streptococci*[182].

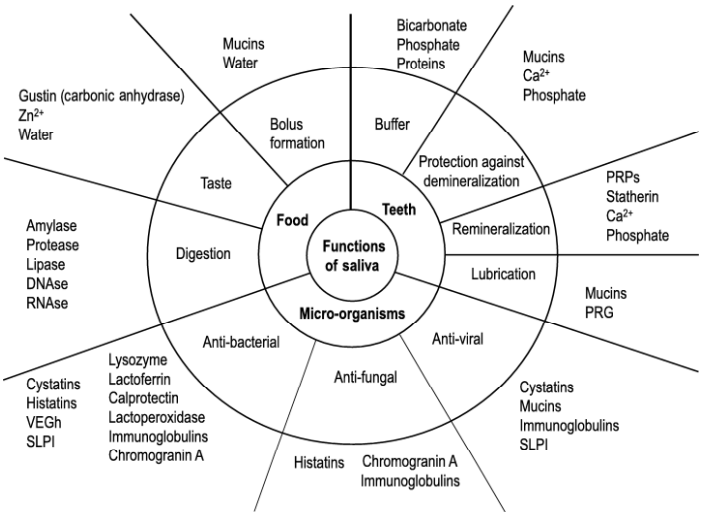


Figure 1.5 Functions of saliva. Proteins found in saliva have various functions including initiating the digestion of food, protecting teeth against demineralisation and antibacterial properties. Adapted from Hirtz *et al.*[180]

The fluids collected from the mouth are known as oral fluids or 'whole saliva', which is distinct from 'saliva', which constitutes the glandular secretions. Oral fluids (or whole saliva) include stimulated and unstimulated saliva, gingival crevicular fluids, oral transudate, pathological exudates, nasal secretions, regurgitated gastric secretions, nonadherent oral bacteria and food debris, and traces of introduced chemicals or medications. Gingival crevicular fluid is secreted from the space between the

teeth and gums and the rate of secretion increases with damage to the gum/tooth interface. Nasal secretions pass into the back of the throat constantly. Gastric contents, particularly after drinking alcohol, will be regurgitated into the mouth. Saliva is produced by three major paired exocrine glands (Figure 1.6): the parotid (P), submandibular (SM) and sublingual (SL), as well as many minor salivary glands. Healthy humans produce around 1-1.5 litres of whole saliva a day, approximately 65% from submandibular, 20% from parotid, 7-8% from sublingual and less than 10% from the minor salivary glands[182]. Unstimulated saliva flow above 0.1 ml/min is considered as normal and the average is 0.3 ml/min[183]. Saliva flow rate can be reduced by dehydration, fear, sleep, or use of some drugs such as atropine and opiates. Production is increased with chemical (gustatory or olfactory), physical (e.g chewing) or psychological stimulation. As with Pavlov's dogs, where the association of a bell with food can make dogs drool, associations linked with food cause greater salivation in humans[184]. With stimulation, the percentage contribution from each of the glands changes, and the percentage contribution from the parotid gland increases from 7-8% to over 50%[185]. The maximum stimulated flow rate is around 7 ml/min, and up to 80-90% of the saliva produced a day may be stimulated[185].

There are three secretory pathways for saliva; the major-regulated pathway, the constitutive-like pathway and the minor-regulated pathway. The major-regulated pathway is regulated primarily by β-adrenergic stimulation. Agonist binding to the β-receptor results in an increase in cAMP in the cell that activates PKA and an increase in exocytosis[186]. The major-regulated pathway is responsible for the high-abundance saliva proteins and provides most of the stimulated saliva during parasympathetic innervations[187]. On the other hand, the constitutive-like and minor-regulated pathways are mainly responsible for the production of unstimulated saliva during basal parasympathetic stimulation[188]. It appears that exosomes and apoptotic blebs are also secreted in saliva, and proteomics have been carried out on these membrane-bound vesicles[187].

In this project, whole saliva was collected, and consequently all the oral fluids were analysed together (Chapter 3). The rationale behind collection of whole saliva relates specifically to application of this work to 'spit tests' in the future; whole saliva is more convenient to collect than individual saliva components (e.g. individual saliva gland secretions). Furthermore, whole saliva collection is more comfortable for the patient and does not require a dentist for sample collection[189]. The main sources of proteins in oral fluid are the salivary glands, but there are also proteins in the other oral fluids mentioned above as well as from microorganisms (mostly bacteria), blood and oral tissues that could be important contributors, especially from people with poor oral health. In the literature, a distinction between saliva and oral fluids is generally not made, and the salivary proteome comprises oral fluids and often contaminants such as blood and bacteria.

Saliva secretions are classified as serous, mucous or mixed. The parotid produces serous secretions, the minor salivary glands produce mucous secretions and the SM and SL have both serous and mucous secreting cells.[185] The composition of saliva secreted is unique to each of the 3 different glands, and it is possible to collect secretions from individual glands by intraoral devices that are

capable of isolating a glandular duct. Because saliva is produced in acinar cells it is isotonic, but it changes to hypotonic on passing through the ductal network. This allows the mucin glycoproteins to hydrate and consequently expand to cover and protect the mouth tissues[190]. This is particularly important during low flow, such as during sleep.

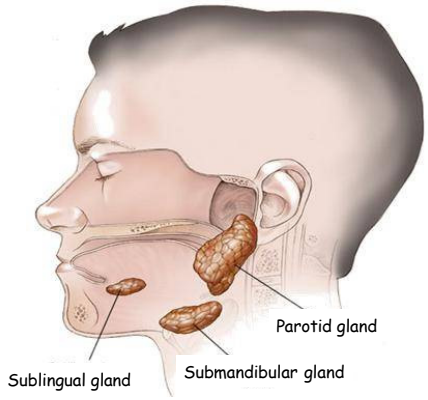


Figure 1.6 Main salivary glands. Saliva is secreted from three main paired exocrine glands: sublingual, submandibular and parotid. The sublingual glands are located below the tongue, the submandibular are located below the lower jaw and the parotid gland is located anterior to the external ear and wraps around the mandibular ramus. Adapted from Wong *et al.*[189]

The SM and SL ductal orifices are close together, so obtaining secretions from individual ducts can be difficult. In some individuals these glands may share common ducts. In order to ensure the purity of SM and SL secretions, the Loo laboratory developed unique biomarkers that can be used to identify pure SM and SL secretions. Cystatin C was found to be unique to SM secretions and MUC5B mucin and Calgranulin B were found to be unique to SL secretions[191].

1.4.2 Salivary proteomics

Salivary proteomics is a relatively new field and less work has been carried out on saliva than on other body fluids, such as blood and urine. The main reason for this is that the concentrations of proteins in saliva are generally lower than in blood but, with the advances in proteomics, the analysis of lower protein concentrations is now possible. There are various research groups investigating the salivary proteome with a view of cataloguing all the proteins present[180, 189, 192-204]. One of the main initiatives is by the National Institute of Dental and Craniofacial Research (NIDCR) to develop tools for disease diagnosis using saliva. In 2003 NIDCR funded three institutions – UCLA, USC and Scripps to catalogue the salivary proteome and to identify all the salivary proteins from the 3 major salivary glands. This initiative has 3 main aims; the first aim involved clinical acquisition of saliva from the 3 major salivary glands (parotid, submandibular, sublingual) from a cohort of human donors. The second aim was proteomic analysis to decipher the complete salivary proteome and its associated complexes from the three major glands. The third was the biocomputational and bioinformatics

cataloguing of the salivary proteome. The work was published in recently and the Salivary Proteome Knowledge Base is available on the internet[205].

A total of 1,116 proteins were identified in the ductal secretions of saliva: 914 in parotid and 917 in SM/SL. 57% were identified in both parotid and SM/SL secretions. The majority of proteins identified are linked to extracellular and secretory classes. The three research groups identified different numbers of proteins; however, 152 parotid and 139 SM/SL proteins were identified by all three research groups (Figure 1.7 and Figure 1.8).[205]

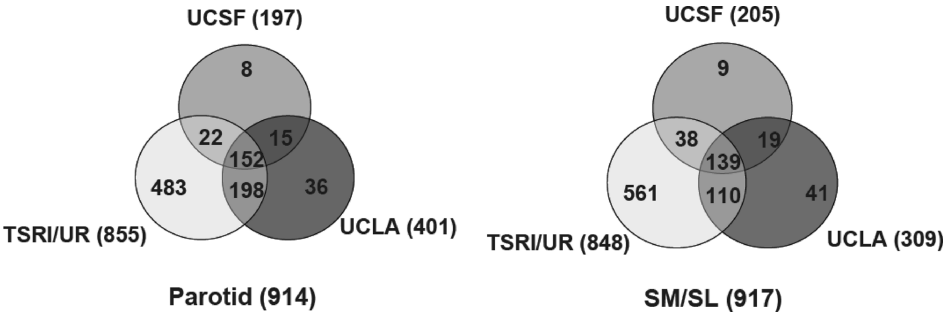


Figure 1.7 Venn diagrams showing the number of proteins by each research group in the NIDCR saliva catalogue initiative. Figure adapted from Denny *et al.*[206] Relatively poor overlap was observed between the three research groups on account of different methodologies used.

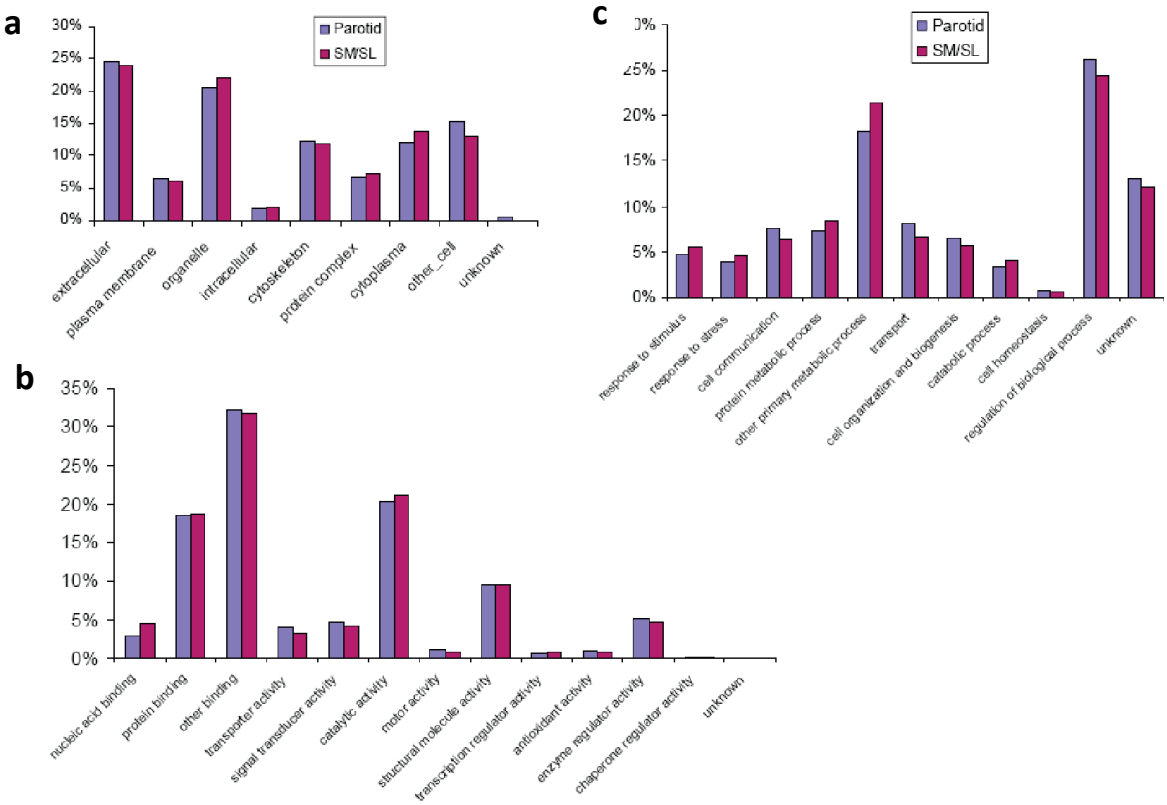


Figure 1.8 Classification of proteins in parotid and SM/SL saliva based on their GO annotation. Figure adapted from Denny *et al.*[206] The functional classifications of parotid and SM/SL are very similar across the different categories: (a) cellular location, (b) molecular function and (c) biological processes.

Many of the proteins identified in saliva (from the three research study) were also present in plasma and tears (Figure 1.9)[205, 207]. 192 out of 657 plasma proteins were identified in saliva and 55% of tear proteins were identified in saliva (259) [189, 208-210].

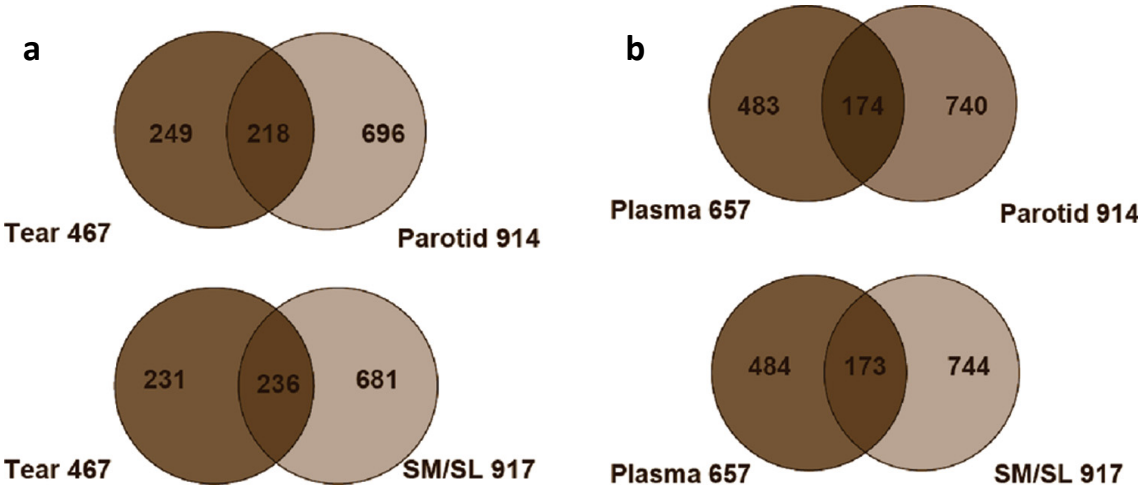


Figure 1.9 Venn diagrams showing the overlap of proteins in parotid and SM/SL saliva with plasma and tear. Many proteins present in plasma and tear are also found in saliva. Figure adapted from Denny *et al.*[206]

Many studies have been performed on the saliva supernatant (Table 1.2). Most of the studies on ductal saliva have focused on the parotid[187]. Some researchers have also focused on subsets such as the salivary glycoproteome[200, 202] but only one laboratory has recently considered the cellular portion of saliva[211].

Table 1.2 Main salivary proteome results of the saliva supernatant. The most comprehensive list of proteins identified in whole saliva is from the Griffin lab where 2340 proteins were identified using a three dimensional peptide fractionation approach[107].

Procedure	Number of protein IDs	Reference
2D SDS-PAGE-LC-MSMS	309	Hu *et al.*[212]
2D LC-MSMS	102	Wilmarth *et al.*[193]
Free-flow electrophoresis-LC-MSMS	437	Xie *et al.*[213]
Various separation LC-MSMS	1166	Denny *et al.*
Capillary IEF-LC-MSMS	1381	Guo *et al.*[192]
3D IEF-LC-MSMS	2340	Bandhakavi *et al.*[107]

The proteome of parotid exosomes was investigated by the Yates group and 491 proteins were observed. 265 of these had previously been reported from parotid gland secretions and 72 had been identified in the proteomic analysis of urinary exosomes. Of the parotid exosome proteome, 43% were assigned to cytosolic proteins according to GO and KEGG pathway analysis, and 26% were integral plasma membrane and associated/peripheral plasma membrane proteins (Figure 1.10).

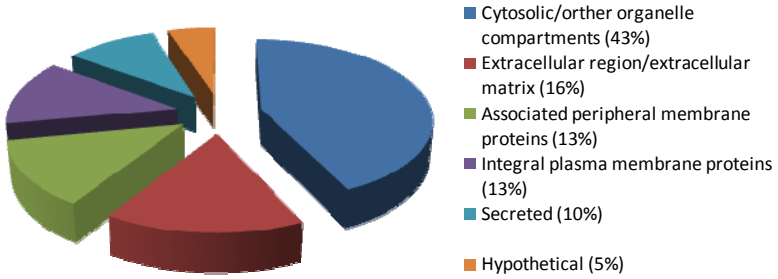


Figure 1.10 Parotid exosome proteins classified by subcellular location. Figure adapted from Gonzalez-Begne *et al.*[187]

Most of the salivary proteome studies reported in the literature have been searched against human databases. Few studies have searched their data against bacterial databases. One example identified 31 bacterial species in saliva supernatant including *Streptococcus mutans*, *Streptococcus pneumoniae*, *Streptococcus avermitilis*, *Porphyromonas gingivalis* and *Helicobacter pylori*[192] and another study identified around 30 bacterial species in the saliva residue[211].

The majority of work to date has focused on cataloguing the salivary proteome. The samples are generally pooled or taken from a single individual, and there does not appear to have been a thorough investigation of the amount of variation between individuals or the effect of circadian rhythm, health status, age, or gender. It is likely that the protein composition of saliva will change quite significantly with the above factors, along with the food consumption of the individual, the amount of physical exercise, and medication. Investigation of the salivary dynamic proteome is important for biomarker studies to determine the 'normal' amount of variation in the quantities of the various proteins.

Although technologies such as MS and other protein separation techniques have been developed to identify and quantify proteins, there is no one universal strategy to identify all proteins in a biological sample. The main problem arises from the dynamic range of protein amounts and the structural complexity. The main steps are the purification and fractionation, identification of proteins and PTMs, followed by quantification.

1.4.1 Salivary microbiome

As a continuation of the Human Genome Project, the NIH sponsored the Human Microbiome Project (HMP) in December 2007. The microorganisms in five different regions of the human body are being investigated: gastrointestinal tract, oral cavity, nose, skin, and female urogenital tract[214]. The number of microorganisms present in the average human is estimated to exceed the number of human somatic and germ cells by a least 10:1. This NIH project aims to produce 1000 sequenced bacterial genomes of species that are associated with humans. Two main goals are to determine whether individuals

share a core microbiome and whether changes in the human microbiome can be correlated to disease states[214].

The Human Oral Microbiome Database (HOMD) was the first to go on-line in March 2008[215] and the others are still in progress. There are currently 600 microorganisms known to live in the mouth[215]. The HOMD aims to categorise each microbe by its 16S rRNA sequence[215]. The mouth has many different environments each of which favours the growth of different suites of microorganisms: e.g. hard surface of the teeth, soft surfaces of the gums, tongue, lining of the mouth, crevicular pockets. The microflora of the mouth is a diverse community and one of the key populations is plaque, a biofilm on the surface of the teeth which has been proposed to be responsible for gingivitis and periodontitis[216].

The study of bacteria comprising plaque is critical for understanding the progression of gingivitis. In order to study the bacterial biofilm growth, a constant depth film fermentor (CDFF) can be used as a model mouth (Figure 1.11)[217]. This apparatus is maintained at body temperature (37°C) and comprises a glass vessel with stainless steel endplates and a rotating turntable contains holes which have plugs containing bovine enamel discs (modelling 'teeth') and which are recessed 50-500 µm to allow a biofilm to develop on the enamel discs. This system can be used to produce single and multispecies biofilms using artificial and/or human saliva[218]. This apparatus can be used to model the effect of different conditions and treatments on the structure and composition of the biofilm. For example, the effect of chlorhexidine[219] and other antimicrobials on biofilms can be used, or gingivitis conditions can be induced/mimicked through the addition of an artificial gingival crevicular fluid formulation[218].

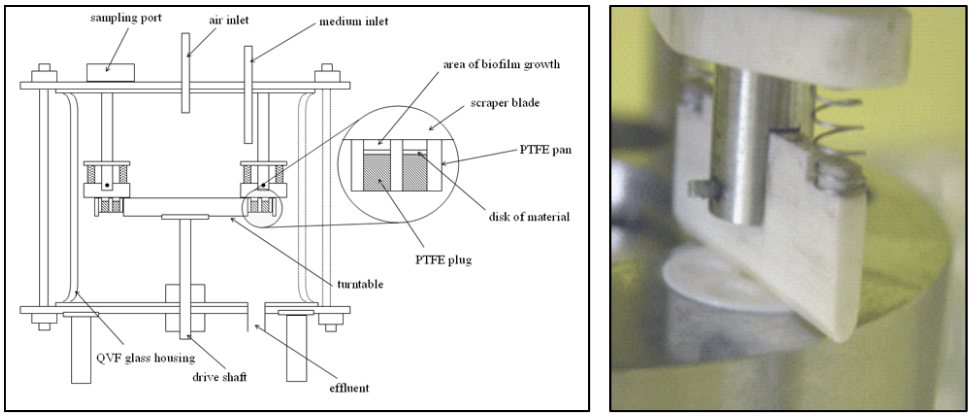


Figure 1.11 Constant Depth Film Fermentor (CDFF) as a mouth model. Adapted from Pratten *et al*. [217]

Whilst many bacteria have been identified in plaque and on other mouth surfaces, very few studies have investigated the presence of bacteria in saliva. One recent study, looking at the cellular residue component of saliva, identified around 30 different bacterial species using proteomics[211]. In many diseases, there is a complex interplay between bacteria and the host response; therefore, it is important to consider as many aspects as possible in the discovery of disease causes and progression, which will translate into biomarkers and ultimately, disease treatment.

1.5 Biomarkers

Biomarkers are biological markers that can be used to indicate a specific state. The term can have various meanings depending on the area in which it is used. In medical terms, it is a substance or pattern of markers that can suggest and/or measure the progression of a disease state. Biomarker discovery is based on various 'omics' technologies including transcriptomics, proteomics, metabonomics, peptidomics, glycomics, phosphoproteomics and lipidomics (Table 1.3)[220]. Transciptomics involves studying the whole set of transcribed RNA products, and many profiling studies have been carried out which suggest that this area has great potential[221,222]. Proteomics, is the study of cellular proteins, is used in this project and has been previously discussed. Metabonomics is concerned with the quantitative metabolic response of multicellular systems to stimuli and relies on both NMR and MS methods. A few metabonomic biomarker studies have been successful in detecting liver cancer using urinary nucleosides as tumour markers[223,224], but relatively few studies have been when compared to proteomic research. Peptidomics investigates all the expressed peptides and their PTMs (post translational modifications). The two main sources of peptidome biomarkers are fragments from protein molecules and cleavage products generated *ex vivo* following blood clotting.[220] Glycomics involves the study of all the glycan molecules in an organism: glycolipids, glycoproteins, lipopolysaccharides, peptidoglycans and proteoglycans. Several studies have shown promise for potential tumour biomarkers[225]. Finally, lipidomics is the characterisation of lipids and their interacting molecules. It has been reported that levels of LPA (lysophopsphatidic acid) and sphingolipids are altered in various cancers[226,227].

Ideally, it would be advantageous to utilise multiple 'omics' technologies to improve biomarker specificity and develop diagnostic tools which may work on a 'personalised' level. There have been several attempts to integrate various classes of biological data, such as linking transciptomics and metabonomics for neuroendocrine cancer[124], and linking genomics, transcriptomics and proteomics for lung cancer biomarker discovery[228]. Integration of omics data, however, is still in its infancy.

Table 1.3 Advantages and disadvantages of various 'omics' technologies. Adapted from Zhang *et al.* [220]

Technique	Advantages	Disadvantages
Transcriptomics	- Well-established technology - Few genes and/or transcripts (ca. 25,000 in humans) relative to proteins	- Tissue materials required
Proteomics	- Suitable for various biological samples	- Many different approaches - More proteins (> 500,000 in humans) relative to transcripts or metabolites
Metabonomics	- Suitable for various biological samples - Fewer metabolites (ca. 10,000 in humans) relative to transcripts or proteins	- Technology in development - Environmental impacts are ignored
Peptidomics	- Low molecular weight	- Proteolysis in *ex vivo* samples complicate the results
Glycomics	- Increased stability and solubility of glycoproteins relative to unmodified proteins	- Difficulty in glycosylation analysis, particularly structure identification
Phosphoproteomics	- Sub-proteome: reduces the number of proteins to be analysed	- Difficulty in the identification of phosphorylated proteins
Lipidomics	- Sub-metabonomics: reduces the number of metabolites that can be measured	- Technology in development

1.5.1 Salivary diagnostics and biomarkers

Proteins are commonly used in clinical assays for disease diagnosis and monitoring. In fact, over 100 protein assays are currently used in clinical testing laboratories[229]. Most of these diagnostic tests current utilise blood samples, while use of other body fluids such as urine, cerebrospinal (CSF), amniotic, peritoneal, pleural and saliva are largely relegated to purely hospital use[229].

Despite the relative dearth of salivary diagnostic tools, saliva is easy to ship and store and it can be obtained at low cost and in sufficient quantities for many types of analysis. Furthermore, saliva collection is non-invasive, which provides a significant advantage over blood collection. This is particularly true in certain populations such as children, the elderly, mentally handicapped and intravenous drug abusers. Unfortunately, saliva contains lower concentrations of protein than blood[230], making analyses which require highly abundant proteins more difficult. However, the relative simplicity and lower dynamic range of the salivary proteome (when compared to blood) also allows for fewer complex protein interactions and would potentially allow for more straightforward, highly sensitive and specific, diagnostic tests to be developed.

Various mouth diseases are known to affect salivary composition including oral cancer[189,231-232] and periodontitis[233]. Periodontitis is a chronic bacterial infection of the tissues surrounding teeth. Symptoms include inflammation, connective tissue breakdown and alveolar bone destruction. There is a need to identify reliable markers for disease activity as a means to assess the effectiveness of treatments and for identifying the 'at risk' individuals who are particularly susceptible to destructive periodontitis[233]. Gingivitis, a milder, early form of periodontitis, is also highly prevalent and having biomarkers to identify it would allow for preventive measures to be taken before the disease worsens. This thesis will investigate potential biomarkers for gingivitis.

Besides oral diseases, saliva tests are currently used in diagnosis of Sjogren's syndrome[234-236], HIV[237-238] and Hepatitis B and C[239] as well as monitoring marijuana, cocaine and alcohol use[240-243]. Saliva can be used to help monitor various systemic conditions such as stress[244], ovarian function[245], ovarian cancer[246,247], tuberculosis, gastric ulcers, liver dysfunction and sarcoidosis[248-251,252,253,254]. Saliva can also be used to test for illegal drugs, or monitor alcohol, steroid and antibody levels[255-256]. Many studies are currently aimed at discovering and validating salivary biomarkers for a range of diseases[230, 236, 257]. Many systemic diseases are known to affect the mouth and the salivary glands which makes searching for related salivary biomarkers a promising endeavour. Examples include diabetes[75], Crohn's disease[258], Lambert-Eaton Myasthenic syndrome (LEMS)[259] and Alzhemier's[260,261].

With the development of proteomics protein biomarker pipelines, there is great potential for increasing numbers of protein biomarkers to be identified, validated and utilised[8]. Saliva offers an ideal body fluid for the diagnosis and monitoring of various diseases, particularly those that affect the mouth, and attempts to identify salivary biomarkers for gingivitis are reported in this thesis.

1.6 Summary

This chapter has given an overview of protein identification and quantitation of proteins using mass spectrometry proteomics. Saliva, in the context of potentially useful body fluid biomarkers was introduced and a brief literature review of the work performed on the salivary proteome was included. The concept of a mass spectrometry based protein activity assay to measure kinase acitivies was outlined (KAYAK). Finally, a short introduction to biomarkers, with emphasis on saliva, was provided. The subsequent chapters of this thesis discuss work involving the investigation of the salivary proteome (Chapter 3), the further development and validation of the KAYAK technique (Chapter 4), and preliminary salivary biomarker discovery for gingivitis using both proteomics as well as KAYAK (Chapter 5).

2 Experimental and method development

This Chapter outlines the experimental procedures used for sample preparation and mass spectrometry both at The University of Edinburgh (UoE) and at Harvard Medical School (HMS). Preparation of saliva for proteomic analysis was different in the two locations and different mass spectrometers were used. The mass spectrometers used at UoE included a quadrupole time-of-flight (QTOF Ultima, Waters), an ion trap (High Capacity Trap, HCT, Bruker) and a Fourier transform ion cyclotron resonance mass spectrometer (FTICR, Bruker). The mass spectrometers at HMS included an ion trap (LTQ, Thermo Fisher) and ion trap-Orbitrap (LTQ Orbitrap, Thermo Fisher).

The procedure for the kinase assay is included along with the optimisation and developments made to the original assay. The procedure for correlation profiling using kinase activity/kinase abundance using anion exchange chromatographic separation is described. This technique is used to identify the kinase responsible for a specific phosphorylation event. As well as using in-solution chromatographic separation, gel separations techniques were investigated and the development and optimisation of in-gel correlation profiling using blue native polyacrylamide gel electrophoresis is described. A description of the molecular biology techniques used is included. These procedures were performed for sample preparation for kinase assays and for validation experiments for putative kinase-substrate pairs following kinase activity/kinase abundance correlation profiling.

2.1 Materials

Unless stated otherwise, all chemicals were obtained from Sigma-Aldrich (St. Louis, MO). Trypsin (mass spectrometry sequencing grade) was obtained from Promega (Madison, USA). Monocolonal anti-HA antibody (F-7, sc-7392) was obtained from Santa Cruz Biotechnologies. Purified human active kinases of PAK3, PAK2, EIF2AK2, PRKACA, MAPK14 and PRKCD were obtained from SignalChem. PBK was obtained from Cell Signaling Technologies and Cdc2/Cyclin B was obtained from Millipore. siRNA for PAK3, TP53RK, CDC2, MAPK14, PBK were obtained from Thermo Fisher. Peptides were synthesised by Cell Signaling Technologies (CST). Briefly, the MultiPep instrument from Intavis Bioanalytical Instruments AG was utilised in a 96 well-plate format using Fmoc chemistry, as described[168]. Unpurified peptides were obtained from CST and then purified individually using semipreparative HPLC on a C18 column and quantified using the spectrophotometric method based on the reaction of free amino groups in amino acids with 2,4,6-trinitrobenzenesulphonic acid[262].

2.2 Sample preparation at UoE (University of Edinburgh)

2.2.1 Saliva preparation

Whole saliva was collected on ice from healthy non-smoking subjects at least 2 h after eating. Saliva was expectorated through a pipette tip into a 15 mL centrifuge tube placed on ice. No stimulation to increase saliva flow was used. Protease inhibitor cocktail (Sigma, 1 μl/mL whole saliva) was added to minimise protein degradation. The saliva was centrifuged at 12,000 rpm at 4°C for 15 min and the residue discarded. Protein concentration was determined using a Bradford assay (Bio-Rad). The proteins in saliva were precipitated overnight using ethanol (5 x volume of saliva) at -20°C. The mixture was centrifuged at 12,000 rpm at 4°C for 15 min, the protein pellet washed with 70% cold ethanol and dried by vacuum centrifugation. The protein pellet was redissolved in 50 mM ammonium bicarbonate (ambic, pH 7.8) containing 6 M urea. Dissolution was aided by sonicating on ice (10 rounds of 20 sec each). The sample was reduced using 10 mM dithiothreitol (DTT) for 45 min at 50°C, alkylated in the dark for 45 min using 50 mM iodoacetamide (IA), and further treated with 50 mM DTT to quench excess IA. The sample was diluted with 50 mM ambic to a final urea concentration of 1M and acetonitrile (ACN) added to 10%. Trypsin (TrypsinGold, mass spectrometry grade, Promega) was added at a 1:20 (w/w) enzyme-to-substrate ratio and incubated overnight at 37°C. After digestion, ACN was removed by vacuum centrifugation and purified using ZipTips according to the manufacturer's instructions. Following vacuum centrifugation the purified peptides were resuspended in 0.1% formic acid (FA) for MS analysis. Percentages are referred as 'by volume' (v/v) unless otherwise specified.

2.2.2 Saliva bacteria preparation

Whole saliva was centrifuged at 1500 rpm at 4°C for 15 min. The supernatant was removed and the residue washed twice with PBS (phosphate buffered saline, pH 7.4). The cell pellet was resuspended in PBS (20 mL) and lysed using a Cell Disruptor Machine (Constant Systems Ltd.) at a pressure of 30 kPa, or using detergents (RIPA buffer containing 1% Tritron X100). The lysed sample was centrifuged at 13,000 rpm at 4°C for 45 minutes and the residue discarded. The supernatant was precipitated overnight using ethanol (5 x volume of sample) at -20°C and then centrifuged at 13,000 rpm, at 4°C for 15 min and the supernatant discarded. The protein pellet was resuspended in 50 mM ambic containing 6 M urea. Dissolution was aided by sonicating on ice (10 rounds of 20 seconds each) and then reduced, alkylated and digested as described in Section 2.2.1. Although very effective, the disadvantage of using the Cell Disruptor Machine for bacterial cell lysis is that a minimum volume of 20 mL is required. For small samples, the cells were lysed using detergents.

2.2.2.1 Culturing of saliva bacteria

LB broth (200 mL) was inoculated with 1 mL of saliva and incubated at 37°C with stirring for 16 hours. The bacteria were collected by centrifuging at 5,000 rpm. The bacterial pellet was washed twice with PBS and then resuspended in PBS. The bacteria were lysed in a Cell Perturbator machine at a pressure of 30 kPa. The lysed bacteria were then processed as described above.

2.2.2.2 Strong cation exchange fractionation (SCX)

Digested salivary bacteria were fractionated by SCX chromatography (Ultimate 3000, Dionex-LC Packings, Hercules, CA) system using a PolySulfoethyl A column (2.6 × 200 mm, 5 μm, 300 Å, Poly LC Inc., Columbia, MD). Peptides were reconstituted in 400 μl of SCX buffer A (10% ACN, 100 mM formic acid, pH 3) and loaded onto the column at 200 μl /min flow rate. An isocratic gradient of SCX buffer A for 20 minutes was used, followed by a 20 minute linear gradient from 0–50% SCX buffer B (100 mM ammonium formate, 10% acetonitrile, pH 6), followed by a 20 minute gradient from 0-100% SCX buffer C (600 mM ammonium acetate, 10% acetonitrile, pH 6) at a constant flow rate of 200 μl /min. Fractions were collected at 1 min intervals and those at the beginning and the end were combined.

2.2.3 Saliva phosphoprotein enrichment

Whole saliva (20 mL) was collected on ice as described in Section 2.2.1. Tris acetate buffer and urea were added to saliva to final concentrations of 50 mM and 6 M respectively. This mixture was centrifuged at 13,000 rpm at 4°C for 20 min to remove debris and filtered through 0.2 μm filters (SpinX, Costar). The protein and peptide double IMAC method used in this study was adapted from Collins *et al.*[137] Fast-flow Sepharose with iminodiacetic acid chelating groups (Amersham Biosciences) was charged with $GaCl_3$. The filtered saliva (20 mL, 12 mg) was incubated with 1 mL of the gallium charged resin with mixing at room temperature for 1 hour. The resin was washed with 50 mM tris acetate containing 6 M urea (10 column volumes) and the phosphoproteins were eluted with 50 mM tris acetate containing 6 M urea, 100 mM EDTA and 100 mM EGTA and concentrated in a spin column (Vivaspin 500,Vivascience). The phosphoproteins were diluted with 25 mM ambic containing 1M urea and 5 % ACN and digested with trypsin (Promega) in a ratio of 1:20 (w/w) trypsin-to-substrate at 37°C overnight and then dried in a vacuum centrifuge. A portion of the digested protein IMAC purified sample (200 μg) was desalted, dried in a vacuum centrifuge, and reconstituted in 200 μL buffer (pH 2.5-3) comprising 34% water, 33% methanol and 33% acetonitrile. Self Pack POROS® 20 MC media (Applied Biosystems) was charged with $GaCl_3$. The peptide mix was then incubated with 200 ul of POROS-Ga slurry for 1 h at room temperature. The resin was then loaded into a spin column and washed with 10 column volumes of buffer comprising 34% water, pH 2.5-3, 33% methanol and 33% acetonitrile. Phosphopeptides were eluted with 1 x 100 μL of 200 mM Na_2HPO_4 and 1 x 100 μl of 100 mM EDTA. The purified phosphopeptides were desalted using Zip tips and reconstituted in 0.1% FA prior to MS analysis.

2.3 Sample preparation at HMS (Harvard Medical School)

2.3.1 Saliva preparation

Whole saliva was collected on ice from healthy non-smoking subjects at least 2 h after eating. Saliva was expectorated through a pipette tip into a 15 mL centrifuge tube placed on ice. No stimulation to increase saliva flow was used. Protease inhibitor cocktail (complete protease inhibitor 1x, Roche) was added to minimise protein degradation. The saliva was centrifuged at 12,000 rpm at 4°C for 15 min and the residue discarded. Protein concentration was determined using a Bradford assay (Pierce). The proteins in saliva were purified using a methanol/chloroform precipitation method[263]. Briefly, for 100 µL of protein sample, 400 µL of MeOH were added. After vortexing and centrifuging (12,000 rpm, 1 min), 100 µL of chloroform was added. After vortexing and centrifuging (12,000 rpm, 1 min), 300 µL of water was added. Phase separation occurred following centrifuging (9,000 rpm, 1 min), and the upper aqueous layer was discarded. 300 µL of MeOH were added to the lower phase and the sample vortexed vigorously. The protein was pelleted by centrifugation for 2 min at 9,000 rpm and the supernatant removed. The protein pellet was washed twice with ice-cold acetone and dissolved in a 50 mM Tris-Cl pH 7.5 buffer containing 8M urea, 50 mM EDTA and 0.005% n-dodecyl β-D-maltoside. The proteins were reduced with 10 mM DTT at 37°C for 20 mins and alkylated with 20 mM iodoacetamide for 45 min in the dark and the reaction quenched with 10 mM DTT. The concentration of urea was diluted to 1 M and the proteins digested with trypsin at a 1:100 (w/w) enzyme to substrate concentration at 37°C for 16 hours. The digestion was stopped using formic acid, (final concentration 5%) and the peptides were purified using StageTips.[264] The purified peptides were reconstituted in a 5% formic acid solution containing 5% ACN prior to MS analysis.

2.3.2 Lysis of human and bacterial cells in saliva

Whole saliva was centrifuged at 1500 rpm at 4°C for 15 min. The supernatant was removed and the residue washed twice with PBS (phosphate buffered saline, pH 7.4). The cell pellet was lysed by bead-beating at 4°C (4 cycles of 30 s, with 60 s breaks in between) in a buffer containing 10 mM K_2HPO_4 pH 7.5, 1 mM EDTA, 5 mM EGTA, 10 mM $MgCl_2$, 50 mM β-glycerophosphate, 0.5% Nonidet P-40, 0.1% Brij 35, 0.1% deoxycholic acid, 1mM sodium orthovanadate, 1mM phenylmethyl-sulfonyl fluoride, 5 µg/mL leupeptin and 5 µg/mL pepstatin A. The protein extract was separated from the beads and the lysate was centrifuged at 10,000 rpm for 10 min to remove insoluble material. The clear supernatant was aliquoted out, snap frozen with liquid nitrogen and then stored at -80°C. Protein concentration was determined by a modified Bradford assay (Pierce).

2.3.3 In-gel protein digestion

Where appropriate, proteins were reduced and alkylated prior to running on a gel. Where this was not desirable (*i.e.* with blue native PAGE where the sample is not denatured), the proteins in the gel were reduced and alkylated prior to digestion. The general procedure for in-gel digestion with reduction and alkylation is given below. The sample was loaded onto a precast bis-tris SDS PAGE gel (mini-gel Invitrogen, 4-12%). The proteins were visualised using colloidal Coomassie. The gel lane was excised into 20 bands, each of which was cut into ca. 1 mm cubes and transferred into 1.5 mL eppendorf tubes. The procedure for in-gel digestion was similar to that described previously[265]. Briefly, the gel cubes were destained using 50 mM ambic/50% ACN, dehydrated using ACN, the disulfide bonds reduced with 25 mM DTT at 56°C for 30 minutes and the free sulfhydryl groups then alkylated with 10 mM IA at 25°C for 45 minutes in the dark and then washed with 50 mM ambic to remove excess IA. The peptides were digested overnight with a solution of 12.5 μg/mL trypsin in 50 mM ambic. Following extraction of the peptides with 50% ACN containing 5% FA, the peptides were dried by vacuum centrifugation, purified using StageTips[264] and frozen at -80°C.

2.4 Mass spectrometry at UoE

2.4.1 Ion trap mass spectrometry

The majority of HPLC (High Performance Liquid Chromatrography) separation was performed using a Dionex 3000 HPLC system (Dionex, UK). Some of the LC-MS/MS runs were carried out using a Proxeon Easy nanoLC system and an Eksigent nano pump LC system, but in those cases the gradient settings were the same. Two different platforms were used: nano ESI and regular ESI.

2.4.1.1 nESI-LC-MS and LC-MSMS

LC-MS/MS of peptide mixtures was performed on a RP HPLC (Ultimate 3000, Dionex, UK) with a High Capacity Trap mass spectrometer (HCT Ultra, Bruker) with a nanoelectrospray interface. The samples were first loaded onto a C18 pepmap precolumn (Dionex, UK) and washed for 3 min with the loading solvent, 0.1% FA. The samples were then loaded onto a PepMap capillary column (75 μm x 15 mm; particle size 3 μm, Dionex, UK) for nano-LC separation at a flow rate of 350 nL/min. The buffers used for the LC were (A) 2% ACN/98% H_2O/0.1% FA and (B) 80% ACN/20% H_2O/0.1% FA. A linear gradient was utilised from 0% B to 50% B in 150 min, and then from 50% to 100% in 20 min. After 10 min at 100% B, the column was re-equilibrated in buffer A for 20 min before the next run. For online MS and MS/MS analyses, a metal tip (Proxeon) with an internal diameter of 50 μm was used. The instrument was externally calibrated using Agilent Tune mix. MSMS data was acquired using data dependent software (HCT Ultra, Esquire Control, Bruker Daltonics). The acquisition parameters for the instrument were as specified in Table 2.1. Different gradient lengths were also used, for example, a 20 minute gradient was often used during label-free optimisation. In this case

gradient was utilised from 0% B to 50% B in 20 min, and to 100% in 2 min. After 10 min at 100% B, the column was re-equilibrated for 10 min before the next run. The method was otherwise the same as above except that compounds for MSMS were actively excluded after the acquisition of 2 spectra for 30 s.

Table 2.1 Mass spectrometry acquisition parameters for High capacity trap, HCT (Bruker) mass spectrometer.

Capillary voltage	1600 V
Plate offset	-500 V
Drying gas	nitrogen
Drying gas flow rate	3 litres/min
Source temperature	150 °C
Aimed ion charge control	150,000
Maximum fill time	200 ms
MS scan range	300-1500 m/z
MS scan rate	8100 m/z /s
MSn scan range	300-1500 m/z
MSn scan rate	26,000 m/z /s
CID minimum intensity	100
CID fragmentation amplitude	1V
Dynamic exclusion	120 s (150 min gradient)
	30 s (150 min gradient)
Charge-state screening	rejects singly charged species
maximum number of MSMS per survey scan	3

2.4.1.2 ESI-LC-MS and LC-MSMS

Similar settings to those specified for nanoESI were used except for the following conditions described. A High Capacity Trap mass spectrometer (HCT Ultra, Bruker) was used with an electrospray interface. The samples were loaded onto a monolithic PS-DVB (polystyrene-divinylbenzene) column (500 μm x 5 cm, Dionex, UK) for LC separation at a flow rate of 20 $\mu Lmin^{-1}$. The buffers used for the LC were (A) 2% ACN/98% H_2O/0.1% FA and (B) 80% ACN/20% H_2O/0.1% FA. A linear gradient was utilised from 0% B to 30% B in 150 min, and to 100% in 20 min. After 10 min at 100% B, the column was re-equilibrated in buffer A for 20 min before the next run. MSMS data was acquired using data dependent software (HCT Ultra, Esquire Control, Bruker Daltonics). The acquisition parameters for the instrument were as specified in Table 2.3 except for the following adjustments: capillary voltage 4000 V; nebuliser 17 psi, drying gas flow rate of 5 litres/min and a source temperature was 310°C.

2.4.2 Fourier transform ion cyclotron resonance mass spectrometry

LC-MS of peptide mixtures was carried out on a reversed phase HPLC (Ultimate 3000, Dionex, UK) with an Apex Ultra Qh Fourier Transform Ion Cyclotron Resonance mass spectrometer equipped with a 12T superconducting magnet (FTICR, 12T, Bruker) and a 6 cm Infinity Cell® (Penning trap). As with the ion trap, two set ups were employed: nESI and standard ESI. nESI was performed using a TriVersa Nanomate (Advion Biosciences, Ithaca, NY) and ESI using an ESI source from Bruker Daltonics. The LC conditions were the same as those described above for the HCT. The mass spectra

were externally calibrated using ES tuning mix (Agilent) and data was acquired using HyStar 3.4 software and mass spectrometry parameters are specified in Table 2.2. Fast fourier transforms were performed using DataAnalysis software (Bruker Daltonics).

Table 2.2 Mass spectrometry acquisition parameters for Fourier transform ion cyclotron resonance, FTICR (Bruker) mass spectrometer.

Capillary voltage	4000 V
Source temperature	150 °C
Ion excitation frequency	48-500 kHz at 100 steps of 25 µs
MS scan range	300 and 3000 m/z
MS scan rate	2 scans/s

2.4.3 Data processing and quantitation

The raw data from the HCT was processed using Bruker DataAnalysis Version 3.4 to generate peak lists in Mascot Generic Format (MGF) for submission to Mascot. A signal to noise threshold of 5, an area threshold of 10, and an intensity threshold of 10 were used. Database searches (NCBI-nr version NCBInr_20071130) were performed using the MASCOT database search engine 2.2.1 (Matrix Science, London, UK). The parameters specified in Mascot included trypsin as the enzyme and up to two missed cleavages, charge states 2^+ and 3^+, mass tolerance in MS mode +/- 1 Da, and for MSMS data a tolerance 0.6 Da. Carbamidomethylation of cysteines was specified as a fixed modification and oxidation of methionine was included as an optional modification. At least 1 unique peptide sequence for each protein hit (required Red Bold in Mascot) and a score greater than the Significance Level in Mascot was required. For label-free quantitation the raw data was converted to mzXML files using CompassXport (Bruker).

2.5 Mass spectrometry at HMS

2.5.1 LTQ Orbitrap

Samples were analysed on various different variations of LTQ/Orbitrap: LTQ, LTQ XL, LTQ Orbitrap, LTQ Orbitrap XL, LTQ Orbitrap Discovery and LTQ Orbitrap Velos (Thermo Fisher). The methods for these instruments were very similar and so just one general method is described. Various different gradient lengths were used depending on the particular experiment. In general, for in-depth proteomic profiling, 2 hour analyses were collected for each fraction.

The samples were analysed on an LTQ Orbitrap mass spectrometer equipped with a Nanospray II electrospray ionisation source (Thermo Fisher, San Jose, CA) coupled to an Agilent 1100 series binary pump (Agilent Technologies, Palo Alto, CA). Dried peptides were reconstituted in water containing 5% FA and 5% ACN and 1-4 µL loaded onto a hand-pulled fused silica microcapillary with an internal diameter of 125 µm, 18 cm long, packed with a C_{18} reversed-phase resin (Magic C18AQ,

particle size 5 µm, pore size 200 Å, Michrom Bioresources, Auburn, CA) using a Famos autosampler (LC Packings, San Francisco, CA). The needle tip had an internal diameter of approximately 5 µm. Once loaded, the peptides were separated across a 45 minute linear gradient of 10-37% solvent B (0.125 % FA in ACN) at a flow rate of 300 nL/min provided across a flow splitter by the HPLC pumps. Buffer A comprised 0.125% FA and 3% ACN in water. Data was collected in a data-dependent mode using the TOP10 strategy where the high resolution MS survey scan was followed by 10 MS/MS scans in the linear ion trap for the top 10 most abundant precursor ions[266] with other mass spectrometry parameters detailed in Table 2.3.

Table 2.3 Mass spectrometry acquisition parameters for LTQ Orbitrap (Thermo Fisher) mass spectrometer.

Capillary voltage	1.8 kV
Capillary temperature	225 °C
Resolution setting	60,000
Automatic gain control for FTMS (AGC)	1,000,000 (3,000,000 on Velos)
Automatic gain control for MSn (AGC)	2000
Ion trap MSn max ion time	150 ms
FTMS Full Max ion time	1 s
Normalised collision energy	30
MS scan range	300-1500 m/z
MSn scan range	300-1500 m/z
Dynamic exclusion	30 s
Charge-state screening	rejects singly and unassigned charged states
Maximum number of MSMS per survey scan	10 (20 on LTQ Orbitrap Velos)

2.5.2 Data processing

The database search parameters and processing parameters varied according to the particular experiment. RAW files were converted to mzXML files using the program ReAdW (http://sashimi.sourceforge.net/software_glossolalia.htmL). MS/MS spectra were searched using the SEQUEST search algorithm (version 27, revision 12)[45] against a concatenated target and decoy human International Protein Index (IPI) protein database (ftp.ebi.ac.uk/pub/databases/IPI/current/)[267] using a precursor mass tolerance of 50 ppm. Search parameters for post translational modifications comprised a static modification of 57 Da on cysteine (carboxyamidomethylation) and dynamic modifications of 16 Da on methionine (oxidation). Protein hits were filtered at the peptide level using Xcorr, ΔXcorr, mass accuracy and peptide length using in-house software to false discovery rates of less than 1% using the target-decoy approach as described previously[268].

2.5.3 Quantitation

Various different quantitation methods were performed depending on the particular experiment including spectral counting, SILAC and isotope-labelled peptides.

Spectral counting was employed for the relative quantitation of proteins present in each of the gel bands. The number of spectra observed per protein (number of spectra observed for all the peptides identified in a particular protein) was used as a rough estimate of protein abundance.

SILAC was used in a pulse-chase type of experiment to determine the rate of degradation of various proteins in HeLa cell nocodazole arrested cell lysate. Peptides were quantified using the Vista program.[269] Filtering was based on the predicted isotopic distribution and a tolerance window of ± 20 ppm from the theoretical mass. The background noise was subtracted from each ion chromatogram and the ratio between the area of the heavy and light peptides was calculated. For accurate quantitation, S/N > 3 was required for both heavy and light peptides.

Absolute quantification of peptide phosphorylation was carried out through comparison of heavy labelled same-sequence phosphopeptides (described in Section 2.7.5).

2.6 Molecular biology

Several molecular biology techniques were employed during this project. Cell culture was used to produce cells for analysis of lysates for kinase activities. Molecular cloning and transfection were used to overexpress several kinases for validation experiments to help identify kinase-substrate pairs.

2.6.1 Cell culture

HeLa cells (cervical cancer) were maintained in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% dialysed fetal bovine serum (FBS) (HyClone). For cell lysis, the media was removed, the cells harvested by scraping, the cell pellet was washed with ice-cold phosphate-buffered saline (PBS) and then lysed with ice-cold lysis buffer (10 mM K_2HPO_4 pH 7.5, 1 mM EDTA, 10 mM $MgCl_2$, 50 mM β-glycerophosphate, 5 mM EGTA, 0.5% Nonidet P-40, 0.1% Brij 35, 0.1% deoxycholic acid, 1mM sodium orthovanadate, 1mM phenylmethyl-sulfonyl fluoride, 5 µg/mL leupeptin and 5 µg/mL pepstatin A). The lysate was centrifuged at 10,000 rpm for 10 min to remove cell debris. The clear supernatant was aliquoted out, snap frozen with liquid nitrogen and then stored at -80°C. Protein concentration was determined by the Bradford assay (Pierce). For the cell cycle studies, HeLa cells were treated with nocodazole (0.2 µg/mL) until 90% of the cells were rounded up for G2/M (second gap/mitosis) arrest. The nocodazole-arrested cells were collected by mitotic shake-off and lysed as described above.

2.6.2 Kinase overexpression

Various kinases were overexpressed using Gateway cloning for KAYAK experiments to determine the effect of kinase overexpression on the phosphorylation of particular kinase substrates (Sections 4.3.1.2 and 4.3.4.4).

2.6.2.1 Background information to Gateway cloning

The Gateway cloning system is based on the site specific recombination of bacteriophage lamda[270-271]. It was commercialised by Invitrogen in the late 1990s and allows the efficient transfer of DNA fragments between plasmids. This is carried out using two proprietary enzyme mixes (BP clonase and LR clonase) and a proprietary set of recombination sequences, termed the 'Gateway Att' sites. There are two main steps: the BP clonase reaction and the LR clonase reaction.

Firstly, the Gateway entry clone (pENTR) is made. Att B1 and att B2 sequences are added to the ends of the gene fragment of interest. The PCR product with flanking att B sites is mixed with BP clonase which inserts the att B sequence into the att P sites in the Gateway Donor vector, forming the Gateway Entry clone with att L sites flanking the gene of interest. Secondly, the Gateway Entry clone is transferred into a Gateway Destination vector using LR clonase. The Destination vector contains attR recombination sequences, promoters and tags. This forms an Expression clone which contains the gene of interest with flanking at B sites.[272] The Gateway cassette is a unique sequence which is cloned into the Destination vector. It contains a chloramphenicol resistance gene. This gene is replaced by the gene of interest during the LR clonase reaction, and hence the desired Destination clone can be selected by culture with chloramphenicol since the Destination clones will not be chloramphenicol resistant. The ORFs for the genes in pDONR223 were recombined into the Gateway destination vector: MSCV-N-Flag-HA-IRES-PURO using λ recombinase. HEK 293 T cells were maintained in DMEM supplemented with 10% FBS. Open reading frames (ORFs) for the kinases of interest were cloned using the Gateway cloning system and transfected into HEK 293T cells. The experimental protocol is described below

2.6.2.2 Experimental procedure for Gateway cloning

A total of 22 kinases were overexpressed using Gateway cloning: AKT1, CDK7, CHEK2, MAPK14, MAPK6, MAPKAPK3, PAK2, PKN1, PLK1, PLK2, PRKAA1, PRKCB1, PRKCD, PRKCG, PRKCH, PRKCI, PRKCZ, PBK, TP53RK, CDC2, PRKAA2, AURKB. The Gateway Entry clones with att L sites flanking the desired genes had previously been made by the Harper laboratory at HMS (pDONR223). The Gateway cloning procedure was carried out for the ORFs of 22 kinase genes of interest. The Gateway Entry clone plasmid in *E. coli* was added to 3 mL of LB Broth containing spectinomycin and incubated overnight at 37°C with mixing. Miniprep DNA (Qiagen) was carried out on each culture according to the manufacturer's instructions.[273] Briefly, the bacteria were pelleted (20 s, 15,000 rpm) and resuspended in 250 μL buffer P1.[273] 250 μL of lysis buffer P2[273] was added and mixed thoroughly. 350 μL of buffer N3[273] was then added to neutralise the basic P2 solution. After centrifugation (30 s, 13,000 rpm), the lysate (supernatant) was transferred to the QIAprep spin

columns. The flow through was discarded after centrifugation (30 s, 13,000 rpm) and the QIAprep spin columns was washed with 750 µL of buffer PE[273]. The DNA was eluted using 50 µL of Buffer EB[273].

The LR clonase reaction to transfer DNA fragments from the entry clones to the destination vectors was carried out according to the manufacturer's instructions[272]. Briefly, 1 µL of the purified plasmid DNA from the entry clone was combined with 1 µL of LR buffer, 1 µL of LR clonase enzyme and 1 µL of phage vector and the recombination reaction was allowed to proceed for 2 h.

The product from the LR recombination reaction was used to transform *E. coli* using heat shock. Briefly, 22 µL of recombinant *E. coli* was added to the reaction mixture from above and placed on ice for 20 min. 225 µL of LB broth containing spectinomycin was added to allow the *E. coli* to divide approximately 3 times. After 1 hour, the *E. coli* were plated on Agar plates and incubated overnight at 37°C. After overnight incubation, 1 colony from each plate was picked out and incubated with LB broth with carbomycin overnight at 37°C with mixing.

DNA was extracted, as described previously using the Qiagen DNA miniprep kit.[273] Part of the DNA was digested using 2 enzymes (HINDIII and BAMH1) to check that the correct protein had been transfected. The digested DNA was run on a 1% agarose/ethidium bromide gel (120V). The DNA plasmids were then transfected into HEK 293T cells using Lipofectamine 2000 according to the manufacturer's instructions[274].

2.6.3 Western blotting

Lysates from the 22 kinase overexpression experiment were run on 4 to 12% SDS/PAGE and transferred onto Protran membranes (Whatman). The membrane was blocked with 5% milk in tris buffered saline with tween (TBST, 10mM tris, pH 7.4, 0.9% w/v NaCl, 0.1% Tween-20 detergent) for 1 h at RT, incubated with a 1:2500 dilution of primary antibody in 0.5% milk for 1 h at RT, washed 4 x with TBST for 15 min each. The washed membrane was then incubated with a 1:10,000 dilution of secondary antibody for 1 h at RT. The membrane was then washed as before. The bands were visualized using ECL solution (SuperSignal West Pico, Thermo Fisher).

2.6.3.1 Kinase knock down using siRNA

HEK 293T cells were cultured in DMEM medium supplemented with 10% FBS. Each of the four siRNAs (siGENOME®, Dharmacon, Thermo Scientific) for each gene were resuspended in RNAase and DNAase-free water (100 µL) and pooled for each gene of interest (1 µL from each siRNA stock). The genes of interest included: PBK, TP53RK, CDC2 and MAPK14. Reverse transfection was performed using 1 µL of the pooled siRNA and RNAiMax (Invitrogen) and Opti-MEM® (Invitrogen) according to the manufacturer's instruction in a 12 well plate. Scrambled siRNA (SiCK) was used as a control. The cells were lysed 48 hours after transfection, as described in Section 2.6.1

2.7 Kinase activity assay and optimisation

In this Section, the KAYAK procedure to profile kinase activities is described along with the new set of substrate peptides that are used in the K60 assay. The terminology Knumber refers to the number of substrate peptides in a particular kinase assay. Various developments that have been made to the original KAYAK assay, allowing greater sample throughput, are discussed. The procedure for correlation profiling using kinase activity and kinase abundance is described using anion exchange chromatrography as a separation technique. The development and optimisation of an in-gel kinase assay and correlation profiling using blue native polyacrylamide gel electrophoresis is described.

2.7.1 KAYAK peptide numbering scheme

To aid data interpretation, a new numbering scheme was devised based on dividing the peptide targets into categories according to the predicted kinase thought to be responsible for their phosphorylation Table 2.4. Previously the peptide codes were based on their location in the synthesis plate[150]. A list of the K60 peptide substrate set with their sequences and new and old numbering key are shown in Table 2.5.

Table 2.4 Category naming system of peptides. The peptides are divided into groups based on the kinase that is thought to be responsible for the phosphorylation of the substrate based on its motif.

Category	Predicted Kinase
AA	PKA
CC	PKC
DD	Acidic
MM	MAPK
PP	Proline directed
RR	RSK
KK	AKT
XX	pY_group_1
YY	pY_group_2
ZZ	pY_group_3
QQ	DNA damage
GG	GSK3
CH	CHEK
OO	Other
BB	Basic
AM	AMPK

Table 2.5 The K60 KAYAK substrate peptide set. The peptide sequence and protein from which it is derived is given. Lowercase s/t/y in the peptide sequences denotes the phosphorylation site. Both the new and old numbering system for peptide numbering is provided.

Peptide	New code for peptide	Sequence	Protein Name
A3	KK1	RPRAAtFPFR	Aktide
A6	PP6	RPGPQsPGSPPFR	Neutrophil cytosol factor 1
A9	MM2	EPLtPSGEAPPFR	EGF receptor
A12	MM3	PSTNSsPVLKPFR	Separase
B2	CC1	KKAsFKAKKPFR	PKC substrate peptide
B4	PP5	IPINGsPRTPPFR	Rb
B6	BB2	PKRKVsSAEGPFR	Non-histone chromosomal protein HMG-14
B11	PP7	LKLSPsPSSRPFR	Lamin-B1
C2	PP4	IPTGTtPQRKPFR	Kinesin-like protein KIF11
C6	BB3	TKRSGsVYEPPFR	Phospholylase b kinase reguratory subunit beta
C10	XX1	AENAEyLRVAPFR	EGF receptor
C11	AA1	NKRRGsVPILPFR	Erythrocyte membrane protein band 4.2
D4	CC9	LKIQAsFRGHPFR	Neurogranin
D7	PP9	NLLPLsPEEFPFR	STAT1
D11	XX2	REVGDyGQLHPFR	Tyrosine-protein kinase JAK2
E1	YY1	EPEGDyEEVLPFR	Hematopoietic lineage cell-specific protein
E4	PP1	LLRGPsWDPFPFR	Heat shock 27 kDa protein
E5	AM1	LKRSLsELEIPFR	Serum response factor
E11	RR2	RKRLIsSVEDPFR	TSC2
F6	RR1	RIRTQsFSLQPFR	Nitric oxide synthase, endothelial
F7	YY2	EPENDyEDVEPFR	Hematopoietic lineage cell-specific protein
F12	YY3	VKRRDyLDLAPFR	Proto-oncogene tyrosine-protein kinase receptor ret
G1	CC3	VLLRPsRRVRPFR	Somatostatin receptor type 3
G2	YY9	ELQDDyEDLLPFR	Band 3 anion transport protein
G10	CC8	KKKRFsFKKSPFR	Myristoylated alanin-rich C-kinase substrate
H2	CC2	WPWQVsLRTRPFR	Plasminogen
H5	WW9	EYDRLyEEYTPFR	PI3K regulatory subunit alpha
H8	ZZ1	IYKNDyYRKRPFR	Proto-oncogene tyrosine-protein kinase ROS
J7	PP10	TKRNStPPPSPFR	B10(position in the first plate)
J10	AA10	NKRRGtVPILPFR	C11(position in the first plate)
K5	BB4	LKRSLtELEIPFR	E5(position in the first plate)
K11	KK2	RIRTQtFSLQPFR	F6(position in the first plate)
L3	AA9	SKRRNtEFEIPFR	G5(position in the first plate)
L12	AM2	AMARAAsAAALPFR	AMPKinase substrate AMARA peptide
M5	BB8	KARSQsGTLDPFR	Tuberin
M10	KK8	RGRSRsAPPNPFR	Bad
N3	PP2	YSRAKsPQPPPFR	Heterogeneous nuclear ribonucleoprotein U
N6	KK9	RKRSLsESSVPFR	Zinc Finger Protein 106
N9	WW10	HTDDGyMPMSPFR	IRS1
N10	QQ2	APVPSGsQSGGRGPFR	Mediator of DNA damage checkpoint protein 1
O4	AA2	KIRRLsASKQPFR	
O7	WW1	MDTEVyESPYPFR	Tyrosine-protein kinase SYK
P3	WW8	YDDDNyDYIVPFR	Dual specificity tyrosine-phosphorylation-regulated kinase 1A
P5	MM1	STSTPTsPGPRTHPFR	Ataxin-2-like protein
P6	DD9	ALDVSAsDDEIARPFR	Acylamino-acid-releasing enzyme
P8	GG2	TASGsSVTsLDGTRPFR	similar to NDRG1 PROTEIN
P9	QQ1	NVKYSSsQPEPRTPFR	CHK1 checkpoint homolog
P10	PP3	EVLRPEtPRPVDIPFR	Caspase-9
P12	GG1	HSsPHQsDDEEEPPFR	GYS1 protein
Q1	CC4	LLATAsFTKSSRPFR	AT-hook-containing transcription factor 1
Q2	OO9	LERNDsWGSFDPFR	Nuclear fragile X mental retardation protein interacting protein 2
Q11	PP11	FEEEEsPKKKPFR	YY peptide1
R8	BB1	RRRslLELH	Polo-like kinase 1 delta
R9	CH1	ALKLVRYPsFvITKKK	chk1tide
R10	CH2	AARLERQDsIFYpKKK	chk2tide
R11	WW2	GGEAIyAApFKK	DBR1
S1	WW3	ETDDyAEIIP	FAK
S2	BB9	RRPsYRKLL	ATF-1
T1	OO2	YDDFKLNSsIVEpKEPA	Nuclear factor NF-kappa-B p100 subunit
T2	OO1	SIEKLNSsLHFlQQE	Calmodulin-regulated spectrin-associated protein 1-like protein 1

2.7.2 Sample preparation for kinase activity analysis

Cell lysates and tissue homogenates were prepared in an ice-cold lysis buffer: 10 mM K_2HPO_4 pH 7.5, 1 mM EDTA, 10 mM $MgCl_2$, 50 mM β-glycerophosphate, 5 mM EGTA, 0.5% Nonidet P-40, 0.1% Brij 35, 0.1% deoxycholic acid, 1mM sodium orthovanadate, 1mM phenylmethyl-sulfonyl fluoride, 5 µg/mL leupeptin and 5 µg/mL pepstatin A. Following lysis/homogenization, the lysate was centrifuged at 10,000 rpm for 10 min at -4°C and the supernatant was snap frozen with liquid nitrogen and stored at -80°C. The number of freeze-thaw cycles was monitored since this affects the kinase activity. Kinase activities were always determined on lysate that had only been frozen once following harvesting. Protein concentration was determined by the modified Bradford assay (Pierce).

2.7.3 KAYAK procedure

Four KAYAK procedures were used in this study; classical, dilute and shoot, 96 well, and in-gel.

2.6.2.1. Classical procedure

Synthetic peptide substrates, each at 5 µM, were incubated with 20 µg samples (*e.g.* cell lysate) and kinase reaction buffer containing Tris-Cl (25 mM, pH 7.5), ATP (5 mM), $MgCl_2$ (7.5 mM), EGTA (0.2 mM), β-glycerophosphate (7.5 mM), Na_3VO_4 (0.1 mM) and DTT (0.1 mM). After incubation for 45 minutes at room temperature, the reaction was quenched with 100 µL of 1% trifluoroacetic acid (TFA). A known amount (5 pmol) of internal standard phosphorylated stable-isotope labeled peptides were spiked into the quenched reaction. The solution was desalted using Sep-Pak C18 50 mg cartridges (Waters) and dried using vacuum centrifugation. The phosphopeptides were enriched using IMAC. In detail, IMAC resin (PHOS-Select iron affinity gel, Sigma) was washed twice with a solution containing 25 mM formic acid and 40% acetonitrile. Washed IMAC beads were diluted 6-fold with 25 mM formic acid containing 40% acetonitrile and 90 µL of this suspension was added to the dried peptides and incubated for 1 hour at room temperature. After incubation the suspension was transferred to a pre-equilibrated StageTip containing a C18 Empore solid phase extraction disk (3M) and the resin washed twice with the resin wash solution (25 mM formic acid, 40% acetonitrile) and once with 0.1% TFA. Phosphopeptides were eluted onto the Empore disks using three washes of potassium phosphate (500 mM, pH 7.0). Bound peptides on the Empore disk were washed first with 0.1% TFA and then 1% formic acid. The desalted purified phosphopeptides were then subsequently eluted with a solution containing 50% acetonitrile and 1% acetic acid. The eluted peptides were dried by vacuum centrifugation and resuspended in 8 µL of 5% formic acid of which 4 µL was then subjected to analysis by mass spectrometry (see Section 2.5.1). In general, a set of 60 substrate peptides was used, termed K60. However, for some optimisation experiments, only a subset of peptides was used. In general, each peptide was mixed to a final concentration of 5 µM unless otherwise specified.

2.7.3.1 KAYAK mass spectrometry analysis

The peptides from the KAYAK assay, dissolved in 5% formic acid, were analysed on an Orbitrap mass spectrometer. Various different Orbitraps were used, depending on availability (LTQ Orbitrap, LTQ Orbitrap XL, LTQ Orbitrap-Discovery). In general, only full MS scans were obtained. If the localisation site of the phospho group on the peptide was not certain, MSMS scans were also collected in the LTQ. The HPLC system and the buffers used have previously been described in Section 2.5.1. A 32 minute gradient was used and data were collected from 400-1,500 m/z full MS with a resolution setting of 60,000, AGC setting of 10^6, ion fill time maximum 1 second.

2.7.4 KAYAK data processing

The amount of substrate phosphorylation was determined by comparing the area of the light phosphorylated substrate peptide with that of the heavy phosphorylated internal standard peptide. The phosphorylated substrate peptides are chemically identical to the heavy internal standard and therefore should have the same chromatographic characteristics and ionisation efficiency. The procedures was as described[151]. Briefly, extracted ion chromatograms were generated from the survey full MS scan with ± 10 ppm mass accuracy around the monoisotopic peak of the most abundant charge state. The extracted ion chromatograms were integrated using Quan browser (Xcalibur 2.0.5, Thermo Fisher). Whilst Quan browser generates the area of the extracted ion chromatograms semi-automatically, each chromatogram required manual inspection to ensure that the correct boundaries had been selected. Where this was not the case, the peak boundaries were manually altered in Quan browser. It was required that heavy and light peptides co-elute perfectly for accurate quantitation. Measurements were excluded where the peak height (extracted ion count) was less than 10^4 or where the peak area was less than 1% of the internal standard (50 fmol). Since a known amount of internal standard was added to the reaction (in general, 5 pmol) prior to purification, the absolute amount of product formed could be determined. The product formed/kinase activity was usually presented with units of fmol/μg/min. The ratio of the light to heavy peptide area was divided by the amount of lysate (usually 20 μg) and the time of the reaction (usually 45 min). For correlation profiling studies, the amount of sample was not considered and the units are expressed as fmol/min. For recombinant kinase assays, the amount formed was not divided by the kinase amount or the reaction time and the units are expressed as fmol.

2.7.5 Development of high throughput KAYAK

The three parts of the KAYAK assay: the sample preparation, analysis on the mass spectrometer, and the data analysis are time consuming. For a standard assay of 12 samples (3 replicates, 36 KAYAK reactions), the sample preparation takes approximately 13 hours. Analysis on the mass spectrometer takes 1 hour per sample, totalling 36 hours. Analysis time using Quanbrowser and manual checking then takes approximately 1 day. Typically, the whole process takes around five days for 36 KAYAK

reactions. Various methods to increase throughput were trialled. The 'classical' KAYAK reaction is carried out in Eppendorf tube format. SepPak purification is performed using individual SepPak cartridges, and IMAC is carried out using StageTips in a centrifuge format.

2.7.5.1 96 well plate format

For large batches of samples, a 96 well format was developed. The KAYAK reaction was carried out in 96 well plates. Following reaction completion, the samples were desalted using SepPak 96 well plate and a vacuum manifold. After SepPak purification the peptides could be dried down by vacuum centrifugation in the 96 well plate. IMAC beads were then added to the 96 well plate and incubated by attaching the 96 well plate to a vortex. The IMAC purification was performed in a 96 well plate using a 96 well C18 Empore (3M) extraction plate. Following IMAC, the purified peptides were dried down by vacuum centrifugation in the 96 well plate, reconstituted, and then the same 96 well plate was put on the mass spectrometer, thereby reducing the time taken transferring samples to autosampler vials. These steps dramatically reduce the sample preparation time. Sample preparation of 96 samples using the classical eppendorf format takes approximately a week whilst the 96 well plate format can be prepared in a day. A few of the steps for the 96 well plate KAYAK required optimisation. During the IMAC purification with the 96 well extraction plate and manifold it was critical to centrifuge the extraction plate to fully remove all the salt, otherwise the salt was not fully removed by solely relying on the vacuum. To reduce cross contamination during use of the vacuum manifold for the SepPak purification elution step, P200 pipette tips were added to each of the 'columns' on the 96 well plate to ensure that the a sample did not contaminate other sample wells - which occurred when pipette tips were not added to the columns.

2.6.2.2. 'Dilute and shoot' method

A significant proportion of the sample preparation time involves the IMAC purification stage of the phosphopeptides. For KAYAK experiments which only required low substrate amounts of 1 µM to 0.1 µM, and/or where lower amounts of lysate were required (for example 6 µg rather than the normal 20 µM), the IMAC purification stage could be excluded, effectively reducing the sample preparation time by half. This method was termed 'dilute and shoot' since, following the Sep-Pak cleanup, the sample only required reconstitution ('dilute') prior to MS analysis ('shoot').

2.7.5.2 Decreasing LC-MS method time

The normal KAYAK reaction with 60 substrate peptides takes approximately 1 hour with a 32 minute gradient. For KAYAK experiments involving only a few peptides, gradient time could be reduced. For particular experiments where only one or two peptides were being quantitated, various parameters were optimised to achieve a very short LC-MS analysis time. A column with a 250 µm ID was used, rather than the standard 125 µm, and packed with only 3 cm of resin (rather than the standard 15 cm) to allow a higher flow rate. To accommodate the higher flow rate, the capillary was set to a higher voltage of 2.5 kV rather than the standard 1.8-2 kV. Initially an isocratic elution was investigated, at

both 30 and 50% B; however, the elution peak was very broad so this approach was discontinued. Instead, a very short gradient was used. Gradient lengths from 0.5 min to 8 min were tried and a length of 2 min was found to give optimum intensity for a single peptide. The LC-MS method for 1 peptide takes in total 7 mins.

KAYAK is in the process of being transformed from the 'classic KAYAK assay'[150] which was time consuming to a much more high throughput analysis where four batches of 96 KAYAK (384) reactions can be prepared in two days (which for the classical method takes two weeks). There is still room for further development and the possibility to move to 384 well plates, in which case the time limiting aspect will switch from sample preparation time to mass spectrometer time. Other technologies to incorporate the phosphopeptide enrichment 'online' could also prove useful in the future. As well as increasing the speed various new substrate peptides were also investigated using peptide libraries (Appendix C).

2.7.6 Kinase activity/kinase abundance correlation profiling

Kinase activity/kinase abundance correlation profiling is a technique recently developed by Dr. Kubota[150] to determine kinase-substrate pairs. Briefly, the sample (cell lysate) is fractionated at the protein level using anion exchange chromatography. For each protein fraction, both the kinase activity and kinase content are determined by KAYAK and shotgun proteomics respectively (Figure 2.1). The correlation profiles between the kinase activity and the kinase spectral count are compared and those with high correlation are possible kinase-substrate pairs.

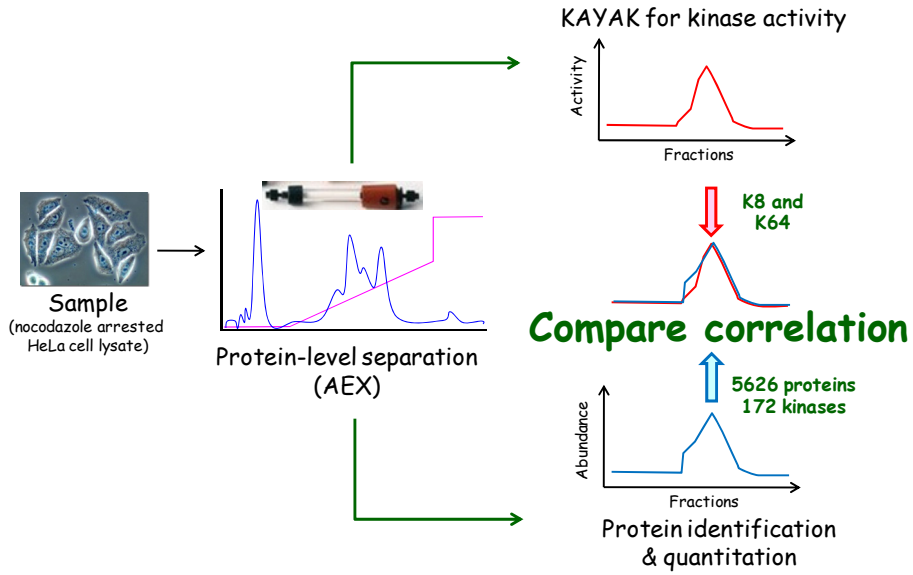


Figure 2.1 KAYAK correlation strategy to identify kinase-substrate pairs. Adapted from Kubota et al.[150] Following protein separation of cell lysate, each fraction is analysed by KAYAK to determine a profile of kinase activities and by LC-MSMS proteomics to obtain profiles of kinase and protein abundance. Comparison of the correlation between the kinase activity profiles for individual peptides and the kinase abundance traces enables identification of possible kinase substrate pairs.

HeLa cell lysate (10 mg), prepared as described in Section 2.6.1, was dialysed in anion AEX buffer: HEPES (20 mM, pH 7.5), NP-40 (0.5%), Brij 35 (0.1%), deoxycholic acid (0.1%), EGTA (1 mM), $MgCl_2$ (5mM), β-glycerophosphate (5mM), Na_3VO_4 (0.1mM), DTT (0.1mM), protease inhibitor cocktail (Complete, Roche Applied Science) and glycerol (20%). The dialysed lysate was centrifuged to remove debris and separated at the protein level using AEX on a monoQ column (5/50 GL, GE Healthcare) with a gradient from 0-0.5 M NaCl in AEX buffer. The lysate was separated into 40 fractions, 500 µL each. Each fraction was aliquoted and analysed by KAYAK and shotgun proteomics. For KAYAK analysis, 30 µL of each 500 µL fraction were used and analysed with a K60 substrate set.

For shotgun proteomics, 100 µL of each fraction were used. For normalisation purposes, 500 fmol BSA were added to 100 µL of each fraction and the proteins were precipitated using the methanol/chloroform method.[263] The precipitates were washed with ice-cold acetone and dissolved in a Tris-Cl pH 7.5 buffer (50 mM) containing urea (8 M), EDTA (50 mM) and n-dodecyl β-D-maltoside (0.005%). The proteins were reduced with DTT (10 mM, 37°C, 20 mins) and alkylated with iodoacetamide (20 mM, 45 min, in the dark).The reaction was quenched with DTT (10 mM). The concentration of urea was diluted to 1 M and the proteins digested with trypsin (5 ng/µL, 37°C, 16 hours). The digestion was stopped using formic acid and the peptides were purified using StageTips.[264]

The relative protein abundance in each of the 42 fractions was determined using spectral counting. The spectral count was normalised to the amount of BSA in each of the fractions as BSA had been added to each of the fractions prior to precipitation. The protein abundance and kinase activity profiles were compared using Pearson correlation[150]. A Pearson correlation coefficient gives a measure of the linear dependence between two variables and has a value from -1 to +1 where '+1' denotes the highest positive correlation, 0 denotes no correlation and '-1' the highest negative correlation. The Pearson correlation coefficient is defined as the covariance of two variables divided by the product of their standard deviation. A Pearson correlation coefficient was determined for each protein (5,626 proteins in total) against each kinase activity profile across the fractions. Fractions that contained less than 5% of the maximum kinase activity were not included in the correlation. The substrate peptides were then ranked with the proteins with the highest ranked Pearson correlation coefficients.

2.7.7 Development of in-gel kinase activity/kinase abundance correlation profiling

Kinase activity/kinase abundance correlation profiling is an approach that can identify kinase-substrate pairs and relies on anion exchange chromatography (AEX) to separate proteins (kinases). The protein profiles are subsequently compared with the kinase activity profiles and the correlation is used to predict kinase-substrate pairs. Whilst this approach is very successful, one of the drawbacks is

that the use of AEX is generally not suited to high-throughput analyses; thus, to enable the kinase activity/kinase abundance correlation profiling method to be more attractive to the wider biological community, a gel-based approach for protein separation would be a more efficient workflow and would allow multiple samples to be separated simultaneously. Proof of concept work was performed to determine if an in-gel kinase activity/kinase abundance correlation profiling was possible. For kinase activity/kinase abundance, it is critical that the kinases remain active and are not denatured, and therefore blue native (BN) polyacrylamide gel electrophoresis (PAGE) was investigated.

BN PAGE was developed by Shägger and von Jagow in 1991[275] and uses Coomassie G-250 (Invitrogen) as a charge-shift molecule which confers a negative charge to the protein without denaturing it. The use of Coomassie results in much higher resolution than native PAGE. Native PAGE is incompatible with nonionic detergents, which are often needed for protein solubility, and also the native PAGE Tris-Glycine system has a high operative pH which may affect pH sensitive proteins. One of the advantages of BN PAGE is that G-250 can reduce protein aggregation: proteins with large regions of hydrophobic surface exposed are less susceptible to aggregation in the presence of G-250 as the dye binds non-specifically to hydrophobic sites, hence converting them to negatively charged sites. Also, the lower pH of the BN PAGE gels (pH 7.5-7.7) is likely to help the proteins retain their structure and activity[276].

2.7.7.1 Proof of concept experiments

Initial proof of concept work firstly involved determining whether protein complexes remained intact following BN PAGE, and secondly whether kinase activity could still be observed. HeLa cell nocodazole arrested lysate was separated by BN PAGE, fractionated into 10 slices and analysed by LC-MSMS, and protein abundance was measured using spectral counting. Around 2,600 proteins were identified (protein FDR 1.5%, peptide FDR 0.15%). To determine whether complexes were still present following BN PAGE, the protein profiles for proteins known to comprise the ribosome and proteasome were plotted for the 10 gel slices. The ribosome comprises two main subunits: the larger 60S subunit (RPL proteins) and the smaller 40S subunit (RPS proteins). The RPL proteins that comprise the large 60S ribosomal subunit were most abundant in the first fraction (red lines, Figure 2.2a), whilst the small 40S subunit proteins were most abundant in the second fraction on the BN PAGE (yellow lines, Figure 2.2a). It was clear that the two ribosomal protein complexes remained intact on BN gel electrophoresis.

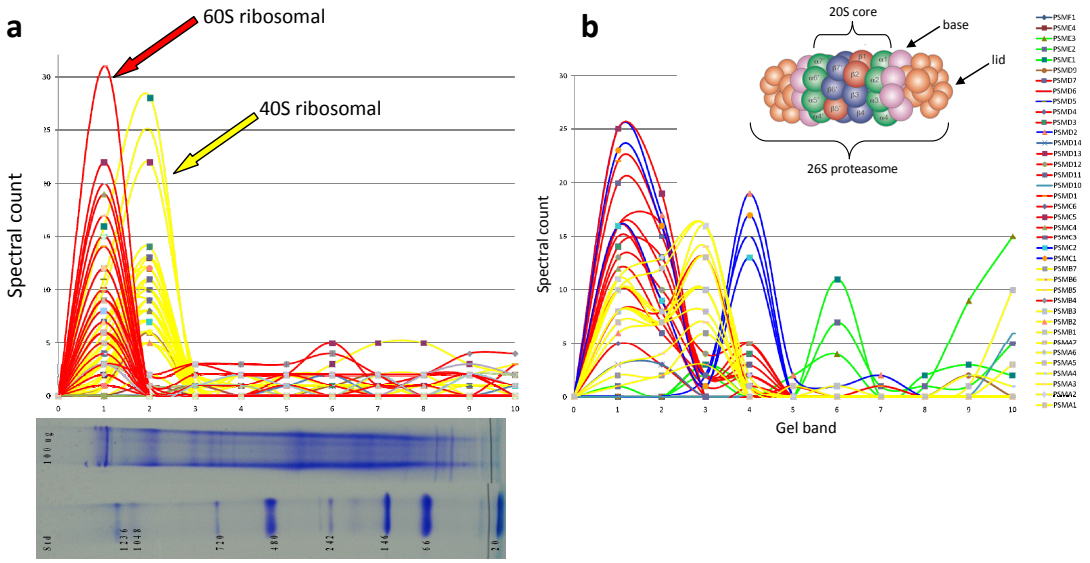


Figure 2.2 Protein profiles from BN PAGE for (a) ribosomal proteins and (b) proteasomal proteins. HeLa cell lysate (100µg) was fractionated into 10 fractions each of which was analysed using LC-MSMS. Protein abundance in each of the fractions was measured using spectral counting. The spectral count of the proteins comprising the ribosome and the proteasome are plotted for each of the gel slices. A picture of the gel is shown on the x axis of the ribosome protein profile graph to illustrate the main protein bands. Proteasome schematic illustrates the base, lid and 20S core, adapted from [277]

The proteasome is the protein machinery responsible for the degradation of misfolded and damaged proteins in the cell. The proteasome consists of a core particle (CP) and a regulatory particle (RP). The CP comprises four stacked rings where the inner two rings comprise seven β subunits and the outer two rings comprise seven α subunits. The regulatory particle comprises 19 subunits and is responsible for unfolding ubiquitin protein conjugates and translocating them into the proteasomal core particle where the proteins are degraded. The abundance profiles of proteins in the core particle (PSMA, PSMB) are similar and peak in the third gel piece (yellow line, Figure 2.2b). Similarly, the abundance profile of proteins in the regulatory particle (PSMC, PSMD) are similar, and peak in the first gel piece (red line, Figure 2.2b). The data confirm that the proteasome complexes have remained intact on the BN gel. Interestingly, the work gave further evidence of the BP1 complex[278] which comprises PSMD5, PSMD2, PSMC2 and PSMC1 and are found to be highly abundant in the 4th fraction of the BN gel (blue line, Figure 2.2b), as described in more detail in Appendix A.

The second proof of concept experiment was to determine whether kinase activity was present in the blue native gel following electrophoresis. To achieve the objective, the KAYAK assay was performed on gel pieces and kinase activity was determined. The presence of kinase activity was observed following BN PAGE, but there was poor resolution and this was evidence for the need for a rigorous optimisation of running conditions.

2.7.7.2 Optimisation of BN PAGE protein separation conditions

The two main problems encountered with BN PAGE were precipitation and streaking. Precipitation occurred in the loading well and also at the 3% to 8% junction (Figure 2.3 a and b). A variety of different conditions were assessed and the BN gel resolution assessed by eye (Table 2.6).

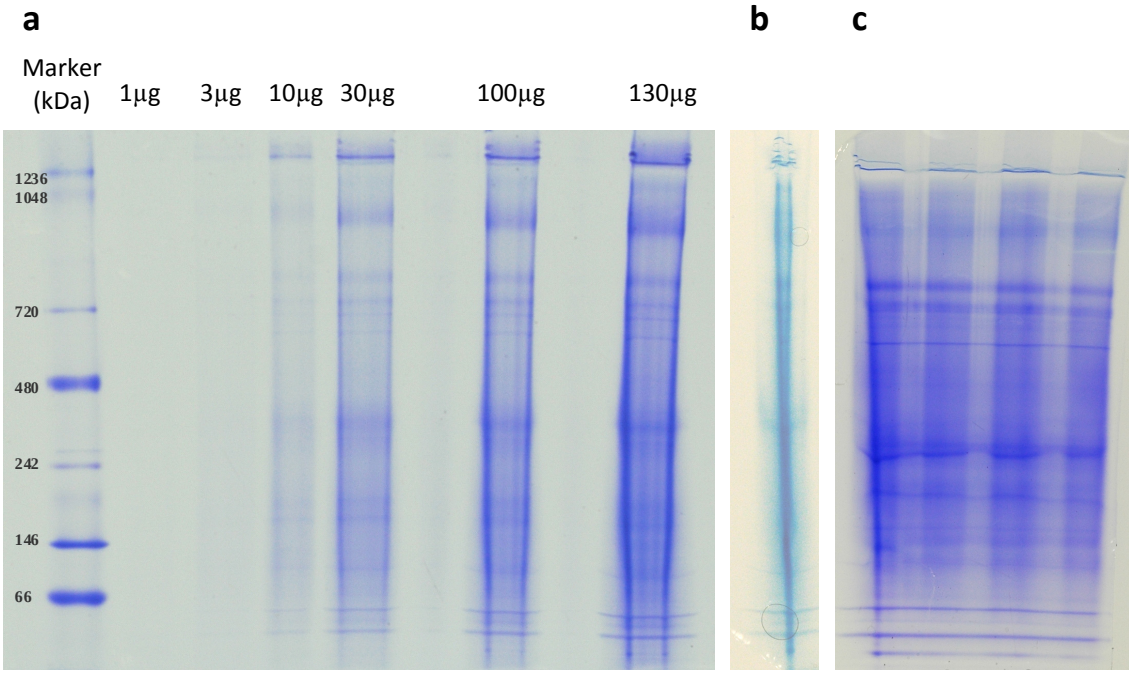


Figure 2.3 HeLa cell lysate loaded onto BN PAGE under various running conditions. (a) BN PAGE with different amounts of protein loaded (1, 3, 10, 100, 130 µg from left to right). (b) BN PAGE displaying bad streaking and precipitation prior to optimisation. (c) BN PAGE run using optimised conditions.

Table 2.6 Description of conditions investigated for optimisation of BN PAGE separation conditions.

Condition	Range	Optimised value
Sample loading	1 – 200 µg	25 – 50 µg
Detergent concentration	0.1-2.5% DDM	0.1%
Cathode running buffer	'light' *versus* 'dark' (Invitrogen)	'light'
Electrophoresis voltage	50-150 mV	50 mV at the start, then 100 mV
% of sample additive (G250, Invitrogen)	0-4x	0.5x
Gel type	'Own cast' *versus* pre-cast commercial (Invitrogen)	Pre-cast gels
Centrifugation of sample	Before and after addition of G250 sample additive	Critical to centrifuge sample following the addition of G250

Streaking of protein bands increased with increasing amounts of protein sample loaded in the wells of the gels (Figure 2.3 a-c). However, streaking was also observed at lower concentrations. The presence and concentration of detergent had a major effect on protein streaking. When no detergent or G-250 sample additive was added, very little streaking was observed; however, the high molecular weight

complexes had very low intensity (low solubility). As the percentage of the detergent was increased, the streaking also increased. It was critical to centrifuge or filter the samples after the addition of the G250, as protein precipitation sometimes occurred following the addition of G250 to the sample. Also, by manually removing the well dividers and loading several lanes, less streaking was observed (Figure 2.3 c). However, it was challenging to remove the well dividers resulting in a completely flat well. Flat wells were necessary for the protein to be loaded evenly to reduce streaking, and any ridges resulted in inconsistent wave-like patterns in the protein bands.

The final optimised protocol was as follows: Sample additive G250 (Invitrogen) was added to the cell lysate (50 µg) which was filtered to remove any precipitation that may have formed. The well dividers in between gel lanes were removed and the protein loaded across several gel lanes. 'Light' cathode running buffer (Invitrogen) containing Coomassie blue dye was used and electrophoresis was run initially at 50 V until the dye front had reached roughly half way and then increased to100 V.

2.7.7.3 Optimisation of in-gel KAYAK assay

Following optimisation of the BN PAGE running conditions, the in-gel KAYAK assay was optimised. The length of time for the KAYAK reaction was evaluated: 11 substrate peptides were used and three gel bands were subjected to an 11 peptide KAYAK assay (K11). The gel was loaded with four samples to allow the KAYAK reaction to proceed for reaction times of 45, 90, 135 and 180 min. The rate of phosphorylation (product formation) was close to linear from 0-90 min (Figure 2.4). For some peptides, at times greater than 90 min the reaction rate decreased and a plateau was observed. To determine whether the reaction rate was linear for shorter reaction times another experiment assessed shorter time periods where sampling occurred at 10, 20, 45 and 90 min (Figure 2.5). The kinase activity was not linear from 0 to 90 minutes, presumably due to a delay in uptake of peptides into the gel piece. This was particularly evident for the peptides displaying high activity, for example B2, C2 and E11 in gel band 2 and E5 in gel band 3 (Figure 2.5). A 90 min assay time was selected as the final assay time since kinase activity was not linear at times over 90 min (plateau was observed), and also non linear kinase activity was observed at shorter reaction times. At the longer reaction times (>90 mins), the substrate concentration could become a limiting factor. The optimised KAYAK in-gel reaction time of 90 minutes was twice as long as the in-solution KAYAK reaction time of 45 minutes.

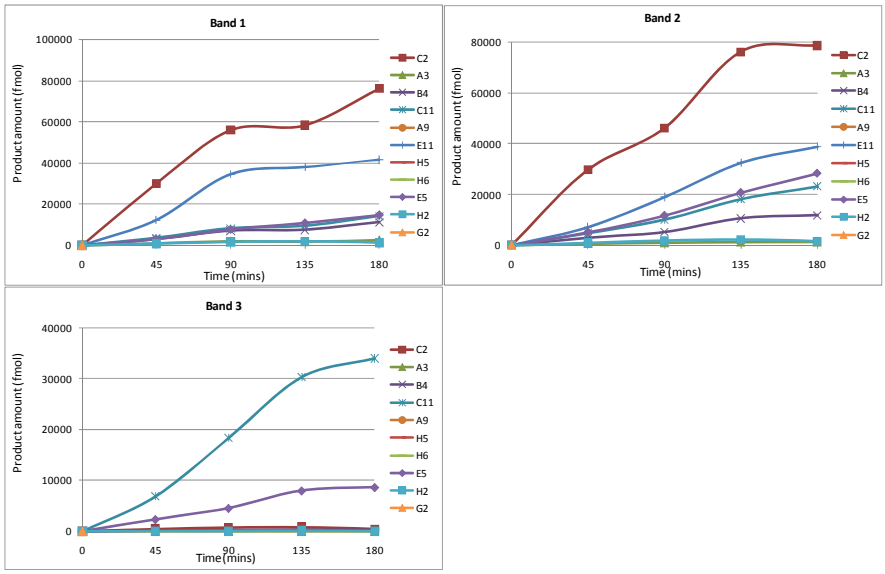


Figure 2.4 Optimisation of in-gel KAYAK reaction time. In-gel KAYAK with 11 substrate peptides was performed on 3 gel bands for 4 different KAYAK reaction times (45, 90, 135 and 190 min).

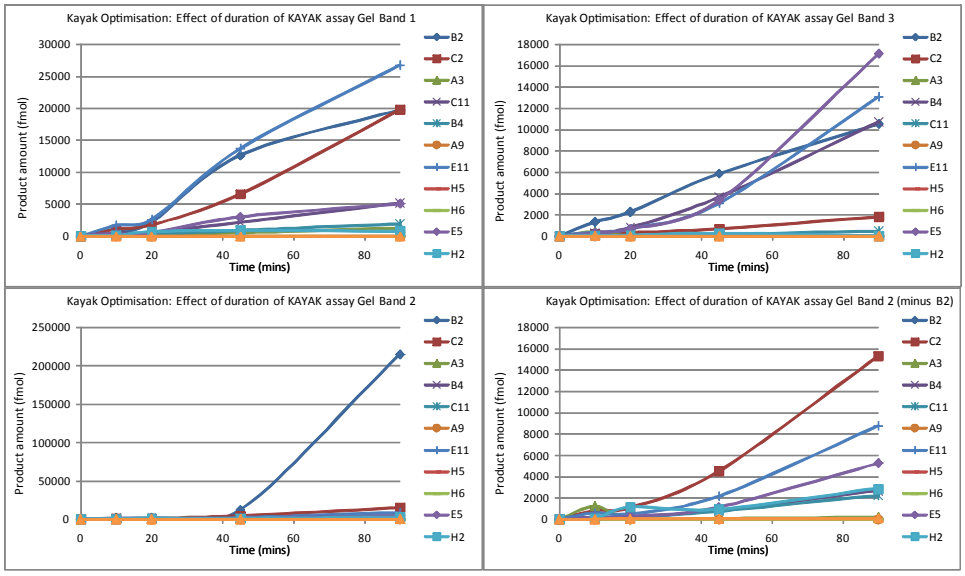


Figure 2.5 Optimisation of in-gel KAYAK reaction time. In-gel KAYAK with 11 substrate peptides was performed on 3 gel bands for 4 different KAYAK reaction times (10, 20, 45 and 90 min).

2.7.7.4 In-gel BN PAGE kinase activity/kinase abundance correlation profiling

Following the successful proof of principle experiments and optimisation of the BN PAGE running conditions and the in-gel KAYAK assay, a correlation profiling experiment was performed with 10 fractions and a K60 substrate set (Figure 2.6).

The lysate (100 µg) was mixed with Coomassie (G-250) and NativePAGE sample buffer (Invitrogen) according to the manufacturer's instructions and loaded onto a precast Bis-Tris nativePAGE gel (mini-gel, 3-12%, Invitrogen) at 4°C. The gel was run at 50 V until the sample had run at least 2 cm into the gel. The voltage was then increased to 100 V for the remainder of the gel electrophoresis (ca. 5 hours). The gel was washed with distilled water and the lane was cut up into 10 equal-sized slices. Each of the slices was further cut up into 1 mm^3 and 100 µL of KAYAK reaction buffer containing 5 µM of K60 substrate peptides was added to the diced gel pieces. The KAYAK reaction was allowed to proceed at room temperature for 90 minutes. All the above steps were performed at 4°C until the KAYAK reaction which took place at room temperature. The kinase reaction was terminated with the addition of a final concentration of 0.75% TFA. The reaction mix was then purified by Sep-Pak C18 columns and analysed by shotgun proteomics for protein correlation profiling as described in Section 2.7.6. After Sep-Pak purification the procedure was the same as that described for the classical KAYAK procedure (Section 2.6.2.1).

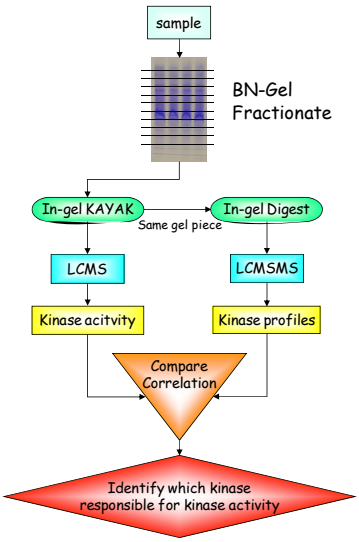


Figure 2.6 In-gel kinase assay correlation profiling workflow using BN PAGE. The sample was fractionated and gel bands were subjected to an in-gel kinase reaction and the resulting phosphorylated substrate peptides were purified using the classical KAYAK procedure. The same gel piece that had been used for the in-gel kinase reaction was then subjected to in-gel trypsin digestion and analysed by LC-MSMS to obtain protein abundance profiles of the proteins present. The protein abundance was determined using spectral counting. The kinase abundance profiles obtained from the spectral counting and the kinase activity profiles obtained from the KAYAK assay were compared using Pearson correlation to identify kinase-substrate pairs.

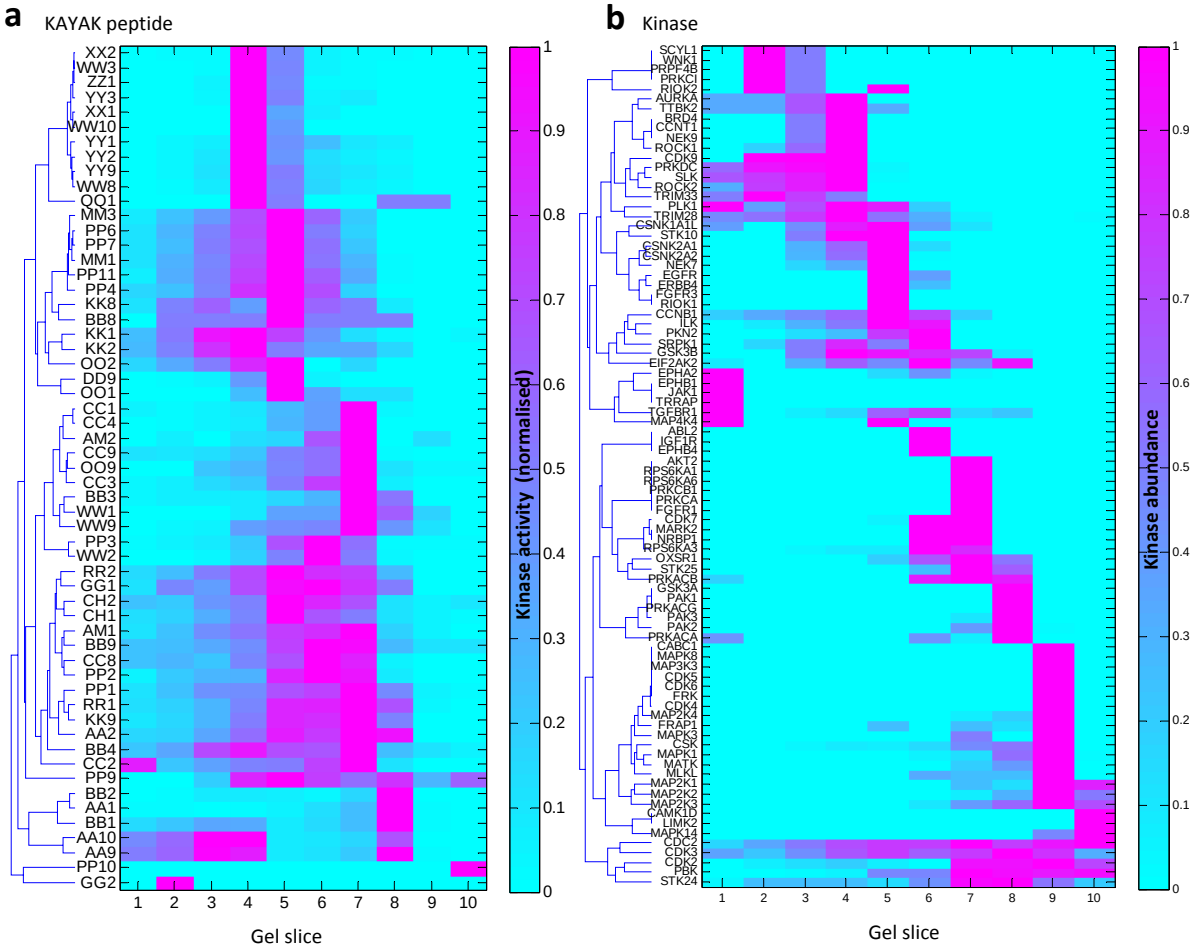
Resolution was a major issue and as a result, it was a challenge to match up the putative kinase-substrate pairs. One of the main problems was due to the low spectral counts of the kinases from the

shotgun LC-MSMS experiments. The kinase activity profiles covered many gel slices, whilst the LC-MSMS often only characterised the kinase in a fewer fractions on account of their low abundance.

There are various ways that the kinase quantitation could be improved, including purification procedures to increase the abundance of kinases, and heavy isotope labelling strategies such as using AQUA for absolute quantitation or dimethylation labelling of alternate fractions. However, the solution sought in this project was to improve the abundance of spectral counts, and hence quantitation accuracy, by the utilisation of an inclusion list for the m/z of the tryptic peptides derived from the kinases. As proof-of-concept for this approach, the m/z of the tryptic peptides for which the kinases had already been identified were used. The list of proof-of-concept candidates comprised peptides from around 100 kinases that had previously been detected with mass spectrometry[150]. This 'target list' approach resulted in the identification of approximately 30% more kinases and the observation of many more spectral counts (Appendix D).

The profiles of the kinase activity (Figure 2.7a) and the kinase abundance (Figure 2.7b) were compared using Pearson correlation (Figure 2.8). Several known kinase-substrate pairs had fairly good correlation coefficients (Figure 2.9) but the profiles spread across several gel bands on account of poor gel resolution.

Figure 2.7 In-gel BN PAGE KAYAK correlation profiling on nocodazole arrested HeLa cell lysate. Gel was fractionated into 10 slices and analysed by KAYAK K60 for its kinase activity and by shotgun LC-MSMS for kinase abundance. (a) Kinase activity profiles across 10 fractions. (b) Kinases abundance profiles across 10 fractions. Hierarchical clustering was performed to group similar responders together. Both the kinase activity (a) and kinase abundance (b) are normalised to the highest value in each row (each gel slice).



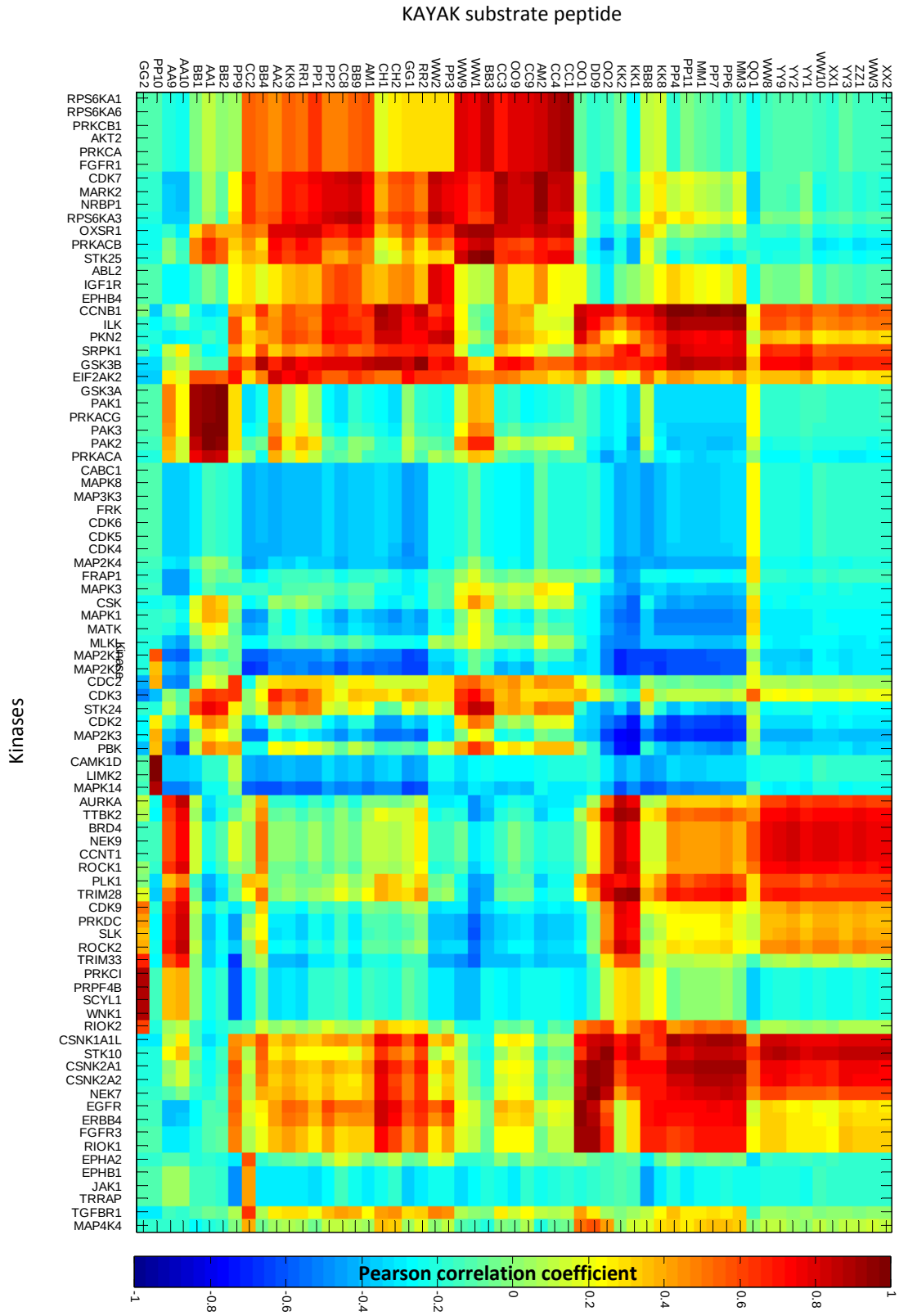


Figure 2.8 In-gel KAYAK correlation profiling using BN PAGE for HeLa cell nocodazole arrested cell lysate. Heat map of Pearson correlation coefficients between the Kinase activity profile (labelled 'KAYAK peptide' on diagram) and Kinase abundance profile labelled 'Kinase' on diagram) across the 10 BN PAGE fractions. Kinase/KAYAK peptides with a high Pearson correlation coefficient (shown as red squares) indicate possible kinase-substrate relationships.

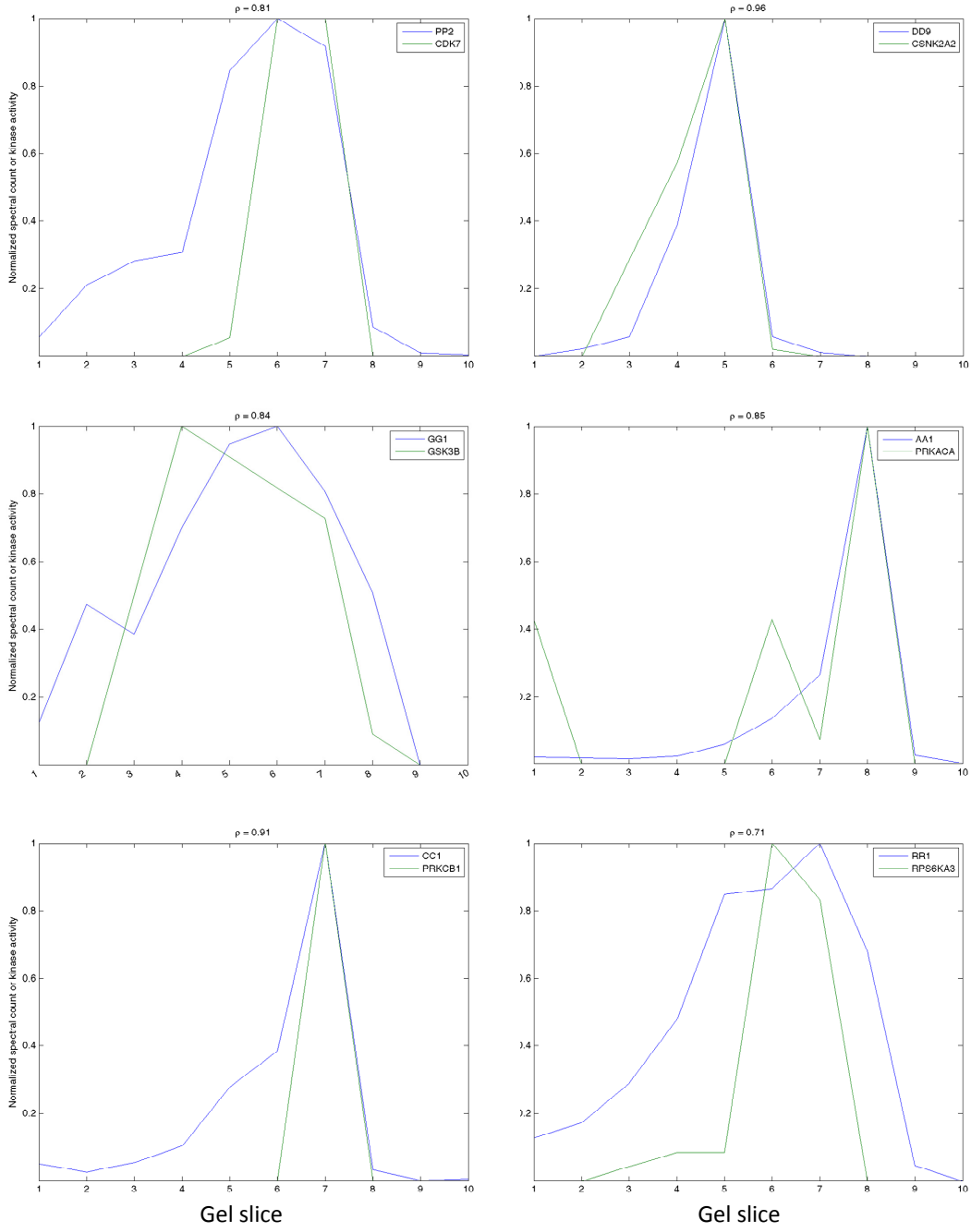


Figure 2.9 Correlation profile plots of selected known kinase-substrate pairs. The normalised spectral count and KAYAK activities are plotted for each BN gel fraction. The Pearson correlation coefficient (ρ) for each kinase-substrate pair is given above each plot.

The low resolution of the BN PAGE along with the limited accuracy of the spectral count kinase quantitation due to the low sample loading capability of the gel reduce the accuracy of the kinase activity/kinase abundance correlation profiling. Whilst the targeted inclusion list greatly improves the quantitation, the spectral counts were still low and the method was ultimately biased towards the kinases on the inclusion list. In theory, a target list comprising all the 600 kinases in the human

genome could be included. A disadvantage of this gel-based method, compared to an in-solution phase separation such as anion exchange chromatography (AEX), is the absence of any internal normalisation. With AEX, a standard protein can be spiked into the separated fractions prior to trypsin digestion, but with the in-gel digestion such normalisation is not feasible.

Proof of concept work for in-gel BN PAGE kinase activity/kinase abundance correlation profiling was performed. The presence of protein complexes and kinase activity was confirmed following BN PAGE and whilst various known kinase-substrate pairs were identified, the procedure suffered greatly from poor BN PAGE resolution. Further kinase activity/kinase abundance correlation profiling experiments performed (described in Chapter 4) used AEX protein separation on account of the greater resolution. However, for future multidimensional correlation profiling experiments, in-gel BN PAGE kinase activity/kinase abundance profiling could prove useful as the second dimension.

A wide variety of techniques have been used throughout this project. Whilst the methods have mostly involved mass spectrometry, a range of molecular biology techniques have also been used and the use of such methods for verification of results obtained by large scale mass spectrometry studies will become increasingly important.

3 Salivary Proteomics

3.1 Introduction

An aim of this PhD was to identify salivary biomarkers for gingivitis. Work at Edinburgh focused on the development of a work-flow for biomarker discovery which could be applied to saliva samples from gingivitis clinical trials. The work in this chapter is divided into two main topics: proteomic analysis of saliva and label-free quantitation. Proteomic analysis of saliva was performed to catalogue the proteins present in different saliva fractions with a view to determine which fraction (SN/res/human/bacterial) would be optimal for subsequent biomarker discovery using label-free proteomics. In parallel, an evaluation and optimisation of label-free quantitation software packages was performed in order to develop a workflow for salivary biomarker discovery. Simple peptide mixtures were initially used to optimise the software parameters prior to analysing saliva samples. The comparison of saliva from gingivitis patients and healthy individuals is covered in Chapter 5 (Salivary biomarkers for gingivitis), whilst this chapter provides the background, methods and preliminary results that set the foundation for the analysis of clinical trial samples.

3.1.1 Proteomics of saliva

A great deal of research, both past and present, has gone into characterising the protein composition of body fluids[210, 279-282]. Analysis of body fluids is currently used for disease diagnosis and their prevalence is expected to increase as robust biomarker discovery and profiling efforts continue. Cellular proteins from numerous tissues are secreted into various body fluids and, by monitoring the relative abundances of these proteins during particular disease states, cellular-level information may be gained that could benefit diagnosis, prognosis, and drug/treatment evaluation[117, 182, 283-284].

To determine the diagnostic potential of saliva, it is critical to first comprehensively identify and characterise the contained proteins. Approximately 20% of salivary proteins are also observed in plasma[205] and, whilst both fluids have comparable disease-linked protein profiles, saliva appears to have higher diagnostic potential for a subset of diseases including cancers, infections, and haematological, gastrointestinal, inflammatory, and respiratory diseases[107]. However, plasma is currently the preferred diagnostic fluid due to its higher protein concentration. Plasma protein concentration is around 50 mg/mL whilst saliva is approximately 1 mg/mL. As with plasma, the large dynamic range of protein abundances in saliva hinders protein identification. Around 10 proteins in saliva account for almost 98% of the salivary proteome, making the analysis of less abundant proteins difficult[285]. Despite these drawbacks, saliva samples are much easier to obtain than plasma, which increases its usefulness as a diagnostic specimen.

Proteomic studies of saliva have largely focused on analysis of the saliva supernatant[211]. Analysis of the residue fraction of saliva has been relatively unexplored, and we believe that protein abundance information of significant value may be found from examination of this fraction. Proteins with very

low abundance in saliva may originate from lysed cells in the mouth. Therefore, analysis of the proteins present in the saliva centrifugation residue (largely unlysed cells and bacteria), and the comparison of these proteins to the species present in the saliva supernatant, can provide an insight into the contribution of proteins from lysed cells to the whole salivary proteome.

Considering different fractions of saliva separately reduces dataset complexity and allows for determination of the fraction which contains the most diagnostic value when comparing samples from healthy individuals and those suffering from gingivitis. In this study, different fractions of saliva were analysed using proteomics, including the supernatant, residue, and the phosphoprotein enriched portion of the saliva supernatant. Although the salivary supernatant has been extensively studied, very few proteomics studies to date have considered these other potentially important fractions[211].

It is unknown whether gingivitis causes a change in bacterial populations or vice versa[286-287]. It may be possible to gain insight into this question by observing changes in the bacterial portion of saliva in sufferers and healthy individuals. Several possibilities exist for the relationship between bacterial and host responses during gingivitis. Changes in bacterial populations (or a specific bacterial species group, species or subspecific variant) in the plaque and/or saliva could lead to changes in the saliva human protein composition and thus contribute to the development of gingivitis. Alternatively, changes in the 'quality' of saliva (for instance, fewer mucins) with environmental factors or ageing could lead to changes in the bacterial population. Interactions of salivary proteins with oral bacteria include aggregation, adherence, cell-killing and inhibition of metabolism, and such interactions would be expected to influence oral biofilm ecology[288]. For example, aggregation of bacteria by salivary proteins causes precipitation and is presumed to help in the clearance of the bacteria from the mouth[289]. Regardless of the mechanism, changes in both the bacterial and human proteome are likely to be observed, and therefore studying both proteomes will determine in which portion the most significant changes occur and potentially in which portion the onset of gingivitis can be first observed.

Another area of salivary proteomics which has been relatively unexplored is phosphoproteomics. There are a few phosphoproteins that have been characterised in saliva and they are known to be important in protecting tooth enamel and integrity. For example, phosphorylation of APRPs (acidic proline rich proteins) maintains the ionic calcium concentration in saliva, and prevents the growth of hydroxyapatite crystals. Phosphorylation of statherin prevents calcium phosphate precipitation[290]. However, to date, a salivary phosphoproteome has not been documented. It is likely that many phosphoproteins are present in gingival crevicular fluid (GCF) due to its contact with bone, which is known to contain extracellular matrix phosphoproteins[291]. Periodontitis is associated with alveolar bone loss from the tooth socket[292] and it is therefore expected that there may be a higher abundance of phosphoproteins observed in saliva which could potentially act as biomarkers for periodontitis. This project will involve the first initial investigation of the salivary phosphoproteome to evaluate whether phosphoprotein enrichment could be of use in the development of a saliva biomarker gingivitis platform.

3.1.2 Salivary protein complexes

Various salivary proteins are known to interact with each other in such a way as to enhance their interaction with oral bacteria[289, 293]. For example, secretory immunoglobulin A (sIgA) has been reported to bind to mucins. Mucins contain hydrophobic domains and sialic acid and sulphate residues which gives them a propensity for homotypic and heterotypic complex formation[294]. There are two major mucins secreted from the SM/SL glands: MG1, a highly glycosylated, high molecular weight mucin; and MG2, a single-glycosylated peptide chain mucin with a low molecular weight[181]. MG1 strongly adsorbs to the surface of the tooth, forming part of the enamel pellicle and protecting the tooth from acid. MG1 is known to form complexes with other proteins such as amylase, proline rich proteins (PRPs), statherin and histatin as well as various bacteria. Interestingly, an MG2-sIgA protein complex was found to bind to the oral bacteria *Pseudomonas aeruginosa* and *Staphylococcus aureus* whilst MG2 alone was found not to bind[294].

A complex interaction between the molecular contents of saliva and bacteria present on the tooth and gum surfaces is likely to play a role in gingivitis. It is therefore of interest to look at the protein complexes that may be present in saliva. Whilst a number of protein complexes have been identified in saliva, to date, there has not been a global proteomics analysis of saliva complexes. Proof-of-concept experiments were performed in this study to investigate the presence of complexes in saliva and to determine their composition. A novel technique combining blue native gels with protein correlation profiling was used to look for putative salivary complexes.

3.1.3 Label-free protein quantitation

The composition of different fractions of saliva were analysed in this thesis. As well as identifying the proteins present in saliva it is critical to quantitate differences in protein abundance for biomarker discovery. A number of strategies are available for protein quantitation and an overview was given in Section 1.2.1. Label-free approaches for quantitation, rather than labelling techniques, were chosen due to their ability to analyse large numbers of samples for biomarker discovery and their relative cost-savings when compared to labelled techniques such as iTRAQ. This section introduces the label-free software packages that were investigated in this thesis.

Three software packages were evaluated in this study: XCMS, MS-Xelerator and Progenesis. This work was performed at the University of Edinburgh. Initially in the laboratory, no quantitative proteomics platforms were set up for either labelled or label-free approaches, and therefore it was necessary to evaluate a variety of software packages. The first label-free software investigated was XCMS. Despite the benefit of being open-source software, at the time of the commencement of this project, XCMS was not able to link LCMS quantitation with the identification LC-MSMS data. Two other software packages were also investigated: MS-Xelerator and Progenesis. It is important to note that both of these software packages were in the beta development stage.

3.1.3.1 XCMS

XCMS was developed at the Scripps Centre for Mass Spectrometry by Smith and coworkers[295] and is available through the Bioconductor bioinformatics project and the Metlin Metabolite Database[98]. Although it was originally designed for LCMS metabolite data it can also be applied to LCMS peptide digests. XCMS comprises three main processing steps for each sample: peak detection, peak matching and retention time alignment. For peak detection, the LCMS filtering algorithm generates EICs (extracted ion chromatograms) in each of the 0.1 m/z bins. Peaks are detected based on the signal-to-noise (S/N) cut-off which can be altered by the user, as well as a limit of 5 peaks per chromatographic slice (see Appendix E.1 for more information). XCMS can be used on both low and high resolution instruments and can be adapted for both centroided and profile data. Following peak detection the peaks are matched between the different samples to enable relative quantitation. Peak matching is a rough way to group the peaks for a more refined retention time alignment to be then performed. Well behaved peak groups are used as temporary standards to calculate a nonlinear retention time profile. A well behaved peak group is defined as a group which does not have many extra or missing peaks and this number may be specified by the user. The retention time deviation profile is then used to warp all the peaks into alignment and peak matching is performed once again to provide a set of reproducible peak groups. The relative intensities of the EIC peaks are then compared between the two groups of samples and a univariate t-test is used to identify peaks which are significantly different between the two samples. Samples are subsequently ranked by ascending p-value.

3.1.3.2 MSX

MS-Xelerator (MSX) is commercially available from MsMetrix (Netherlands)[296]. MSX comprises four modules: Browser, MPeaks, IPeaks and MSCompare. MPeaks can compare 2 different samples and alignment is based on a peak based algorithm. MSCompare allows the comparison of a series of LCMS samples and chromatographic alignment is a feature-based, rather than peak-based, algorithm. Currently LCMS and LC-MSMS linking capability is only possible in the MPeaks module. Future versions of the MSCompare module will implement the peak based algorithm and LCMS and LC-MSMS linking. Pre-processing algorithms include smoothing, despiking and baseline correction, and alignment algorithms include offset shifting, correlation optimized warping (COW), and reference peak warping. Normalisation of the total area of specific peak area is possible. For ratio calculation, the intensity (peak height) of the EIC or the peak area can be used. For complicated samples, it is generally advisable to use the peak height since this appears to be less susceptible to overlapping peaks and misalignment (see Appendix E.2 for more information on MSX)[296].

3.1.3.3 Progenesis

Progenesis LC-MS is commercially available from Nonlinear Dynamics (UK)[297]. The LCMS chromatographic alignment is based on algorithms previously developed for alignment of spots in 2D gels (SameSpots) and involves the addition of alignment vectors. Progenesis allows for manual editing of the peaks picked and to adjust the area over which the integration is calculated. Multiple options exist for viewing the data, either in 2D or 3D, and peptides are ranked by ANOVA p-value.

Progenesis uses a normalisation strategy that is based on the assumption that the abundance of at least 50% of the peptides in two given samples is unchanged. The software uses these peptides, for which abundance is unchanged, for normalisation, allowing many data points to be used for normalisation and ignoring outliers. Progenesis has the potential to link LCMS and LC-MSMS data, although this was still under development at the time experiments were performed (see Appendix E.3 for more information).

These three software packages were optimised and evaluated using a range of sample types of varying complexity including standard peptides, BSA digest and saliva spiked with internal standards.

3.2 Experimental Procedures

Most of the saliva preparation and mass spectrometric analysis procedures are detailed in the Experimental Procedures (Chapter 2). A brief overview is given below.

3.2.1 Saliva preparation

Saliva supernatant was prepared and collected as described in Sections 2.2.1 and 2.2.2 at Edinburgh and HMS respectively. Saliva residue was prepared as described in Sections 2.2.2 and 2.3.2. Saliva was analysed using in-solution digestion (Section 2.2.1), SCX (Section 2.2.2.2) and GeLC-MSMS (Section 2.3.3), and analysed on various mass spectrometers including HCT (Section 2.4.1.1) and LTQ Orbitrap (Section 2.5.1). The phosphoprotein portion of saliva was purified using both protein IMAC and peptide IMAC (Section 2.2.3). The bacterial portion of saliva was prepared without culturing (Sections 2.2.2 and 2.3.2) and with culturing (Section 2.2.2.1). The bacterial biofilm was obtained from Dr. David Spratt from an experimental mouth set up at the Eastman Dental Institute, as described[298], and processed in the same way as the bacterial portion of saliva (Section 2.2.2). Saliva (20 μg) was run on BN gel with the optimised conditions specified in Section 2.7.7.2 and analysed using LC-MSMS protein correlation profiling with quantitation using spectral counting (section 2.7.7.4).

3.2.2 Mass spectrometry and data analysis

Mass spectrometry LC-MSMS was performed using an ion trap (HCT) (Section 2.4.1.1) and LTQ-Orbitrap (Section 2.5.1) and the data processed and searched using Mascot (Section 2.4.3) or Sequest (Section 2.5.3). For protein annotation analysis, protein lists were filtered at a 1% FDR at the protein level and analysed for relative enrichment in MetaCore™ Analytical suite version 4.5 (GeneGo, Inc, St. Joseph, MI). MetaCore™ is a web-based computational platform designed for the analysis of high throughput experimental omics data. Enrichment analysis was carried out across various functional ontologies including Gene Ontology (GO) biological process (BP), GO cellular component (CC), GO molecular function (MF). The results were ranked by p-value which was calculated using a hypergeometric distribution as described[299], where the p-value represents the probability of a mapping event arising by chance given the number of genes in the list of interest relative to all the proteins in the background protein list (whole proteome) for a particular ontology. Further protein function annotation using GO was performed using the Proteome Bioknowledge® Library (BKL) Retriever version 2.4.2[300]. Comparison of the number of proteins assigned to each GO category for both saliva supernatant and residue was carried out. Only the major categories with a p-value < 0.05 were included in the analysis.

3.2.3 Optimisation and comparison of label-free software packages

Samples used for optimisation of the label-free software packages included a mix of standard peptides (bradykinin, enkephalin, angiotensin 2, glu-fibrinopeptide and melittin, each at 5pmol/µl) (Sigma Aldrich), BSA digest and 'wake-up' (WU) and 'after toothbrushing' (T) saliva samples. WU saliva was collected immediately after waking up in the morning after at least 6 hours sleep. No liquid or food was consumed prior to producing the saliva sample. T saliva was collected 30 minutes after teeth had been brushed. No food or liquid was consumed in the 30 minutes between toothbrushing and expectoration of saliva. Standard peptides were injected into the mass spectrometer at 150, 300 and 600 pmol. BSA digest was injected into the mass spectrometer with amounts from 100 fmol to 5 pmol. Peptides and BSA digest were spiked into saliva samples at 150 and 300 fmol.

For optimisation and comparison of the software packages with different samples, triplicate LCMS analyses were performed for each sample. LCMS analyses were acquired using an ion trap (HCT, Bruker) and an FTICR (12T, Bruker) with 20 and 150 minute gradients on a Dionex 3000 HPLC system (Sections 2.4.1.2 and 2.4.2). ESI with a flow rate of 20µl/min was used rather than nESI as a more stable spray was obtained with the higher flow rates. For label-free analysis using the EIC, LCMS data was processed manually using Bruker software (DataAnalysis), and with the three label-free software packages: MS-Xelerator, Progenesis and XCMS. Analysis using Progenesis was mainly performed by the company Nonlinear Dynamics since, at the time, we only had access to the earlier version of their software which was not suitable for low resolution data. A variety of settings were investigated in attempt to optimise performance (Table 3.1).

Table 3.1 Range of optimisation settings used for the preprocessing and alignment of chromatograms in the label-free software packages MSX and XCMS.

	MSX	**XCMS**
Preprocessing	Despiking: 0-5 Smoothing: 0-10 scans Baseline corrections: 0-15	
Alignment	RPW: EIC/BPC COW: (5-30)/(30-100)	Bin, binlin, intlin, cwave Missing value 0-10 Gaussian / fwhm: 10-280
Biomarker map variables	Sensitivity: 0.1-5% Minimum peak width: 0.1-2.6 min Peak area/height for quantitation	

3.3 Results and Discussion

Work was performed on cataloguing the salivary proteome for both the saliva supernatant and residue considering proteins of human and bacterial origin and is discussed in the first part of this chapter. The phosphoproteome of the saliva supernatant was also investigated. In parallel, three label-free software packages were evaluated and optimised with standard mixtures and then used to compare saliva samples and this is discussed in the second part of this chapter (Section 3.3.4).

3.3.1 Proteomics of saliva

This section is divided into the analysis of the human proteins in the saliva supernatant and residue, a short study analysing the phosphoproteome of the human proteins in the saliva supernatant, and an investigation of the bacterial proteins in saliva. In addition to cataloguing the proteins identified in saliva, gene enrichment analysis of the proteins in the different saliva fractions was performed to determine categories that were significantly overrepresented in saliva compared to the human proteome.

3.3.1.1 Proteomics of human saliva in the residue and supernatant

Saliva supernatant was digested in solution and analysed by LC-MSMS on the HCT (Bruker) ion trap mass spectrometer. Around 60 proteins were identified in one LC-MSMS run. The goal was to identify as many proteins as possible in one LC-MSMS run. One LC-MSMS run was preferred due to ease of label-free quantitation and the expectation of large numbers of clinical samples. Considering over 1000 proteins have been identified in saliva[206], the number of proteins identified here was very low. The presence of a few very abundant proteins in saliva, such as amylases and immunoglobulins, dominate the protein mixture and, without sample prefractionation, mask lower abundance proteins.

To identify more proteins in saliva it was necessary to fractionate the saliva. Saliva was fractionated by SDS PAGE into 20 fractions and analysed by LC-MSMS on an LTQ-Orbitrap. A total of 834 proteins were identified with a FDR (false discovery rate) <1% at the protein level, where 4,657 unique peptides and 30,670 total peptides were identified. This was lower than the number of proteins identified by the NIDCR salivary proteome consortium. However, the NIDCR study involved a very large number of fractionations and the combined efforts of 3 research groups[205].

Whilst many research groups have investigated the proteomics of the soluble (supernatant) fraction of the saliva, very few have considered the saliva residue (that is obtained following saliva centrifugation) using large scale proteomics. One of the main symptoms of gingivitis is inflammation of gingival tissue, and therefore it might be expected that the cells that are sloughed off into saliva from the gingiva may contain markers for gingivitis. It is thought that the majority of these cells originate from the oral epithelium[211].

For the same saliva sample for which the supernatant had been analysed, the corresponding saliva residue was analysed by GeLC-MSMS. The sample was fractionated into 20 gel slices and analysed on an LTQ-Orbitrap. A total of 1,426 proteins were identified with a FDR <1 % at the protein level, where 7,865 unique peptides and 39,751 total peptides were identified. A greater number of proteins were identified in the residue portion of saliva on account of the lower dynamic range of protein abundance compared to the saliva supernatant.

The proteins identified in the residue and the supernatant were compared. The overlap between the supernatant and the residue was very high: 75% of the proteins identified in the supernatant were also identified in the residue and 44% of the proteins identified in the residue overlapped with those identified in the supernatant (Figure 3.1). The large overlap implies that a significant portion of the proteins in saliva originate from lysed epithelial or gingival cells from the inside of the mouth.

Whilst there could be contamination of the saliva residue by salivary proteins from the supernatant, which is more common for 'sticky' proteins such as mucins, the salivary residue pellet was washed with PBS to reduce such contamination. Therefore, it is unlikely that a large proportion of the supernatant proteins in the residue fraction are due to contamination.

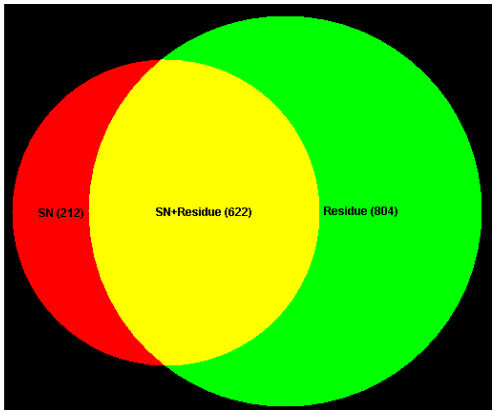


Figure 3.1 Overlap of proteins observed in the salivary supernatant (SN) and reside. Analysis was performed using GeLC-MSMS. SDS PAGE was performed on 20µg of saliva SN and residue respectively and each gel was fractioned into 20 bands, digested in-gel with trypsin and analysed on an LTQ Orbitrap. Of the 834 proteins identified in the saliva supernatant (FDR < 1% at the protein level), 622 of these were also observed in the analysis of the residue portion of saliva. Of the 1,426 proteins identified in the saliva reside, 804 of these were specific to the residue. Drawn using the Convex Venn-3 applet[301].

The spectral count (SC) distribution of co-occurring proteins which were identified in both fractions (residue and SN) was subsequently compared to the SC distribution of proteins in the individual fractions (the SN and residue fractions separately). This was performed to determine whether the co-occurring proteins had low or high spectral count with respect to their counterparts in the supernatant or residue (Figure 3.2a and b). If the large overlap observed between the supernatant and residue proteins was due to contamination during sample preparation, the majority of co-occurring proteins would be expected to be found with low spectral counts in the corresponding sample. However, since

the spectral count distribution of co-occurring proteins in the supernatant and residue was fairly evenly distributed, this would suggest that the overlap is not due to contamination during sample preparation.

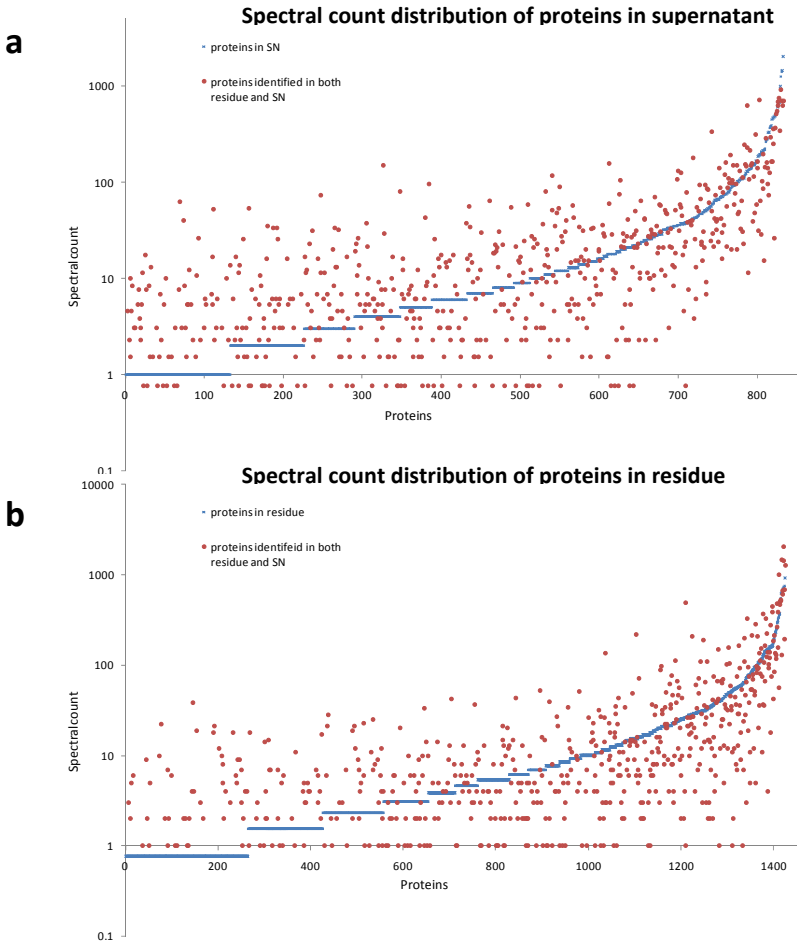


Figure 3.2 Comparison of the spectral count abundance profile of proteins observed in both samples (residue and SN) against all the proteins observed in the individual samples (SN and residue separately). (a) Spectral count distribution of proteins identified in both samples (shown in red) and SN protein (shown in blue). (b) Spectral count distribution of proteins identified in both samples (shown in red) and residue protein (shown in blue). The spectral count of all the SN and residue proteins was normalised to the total spectral count of the total SC of all proteins identified in the SN and residue respectively.

Whilst many studies dedicated to profiling the salivary proteome have concentrated on extensive fractionation and pretreatment of the sample to remove the most abundant proteins in order to identify large numbers of salivary proteins, the approach here was to develop a simple, reproducible method whereby many samples could be processed and compared. One of the disadvantages of removing abundant proteins is reduced reproducibility. A promising new approach using Proteominer beads (Biorad) has been shown to identify a very large number of saliva proteins without the need for removal of abundant proteins[107]. For future studies, the reproducibility of data obtained using these beads will be assessed and this could greatly help in the analysis of salivary proteins.

3.3.1.2 Protein annotation analysis of human saliva

In the allocation of proteins to different gene ontology (GO) categories, the greatest number of proteins for both the saliva supernatant and residue were allocated to the cellular process category for GO biological process (Figure 3.3). For GO cellular component, the greatest number of proteins was assigned to the cytoplasm, and for GO molecular function, binding had the highest number of protein hits. The distribution of the protein profiles across the GO categories was fairly similar for both the saliva supernatant and residue. The largest difference is the greater relative abundance of extracellular proteins in the saliva supernatant compared to the saliva residue.

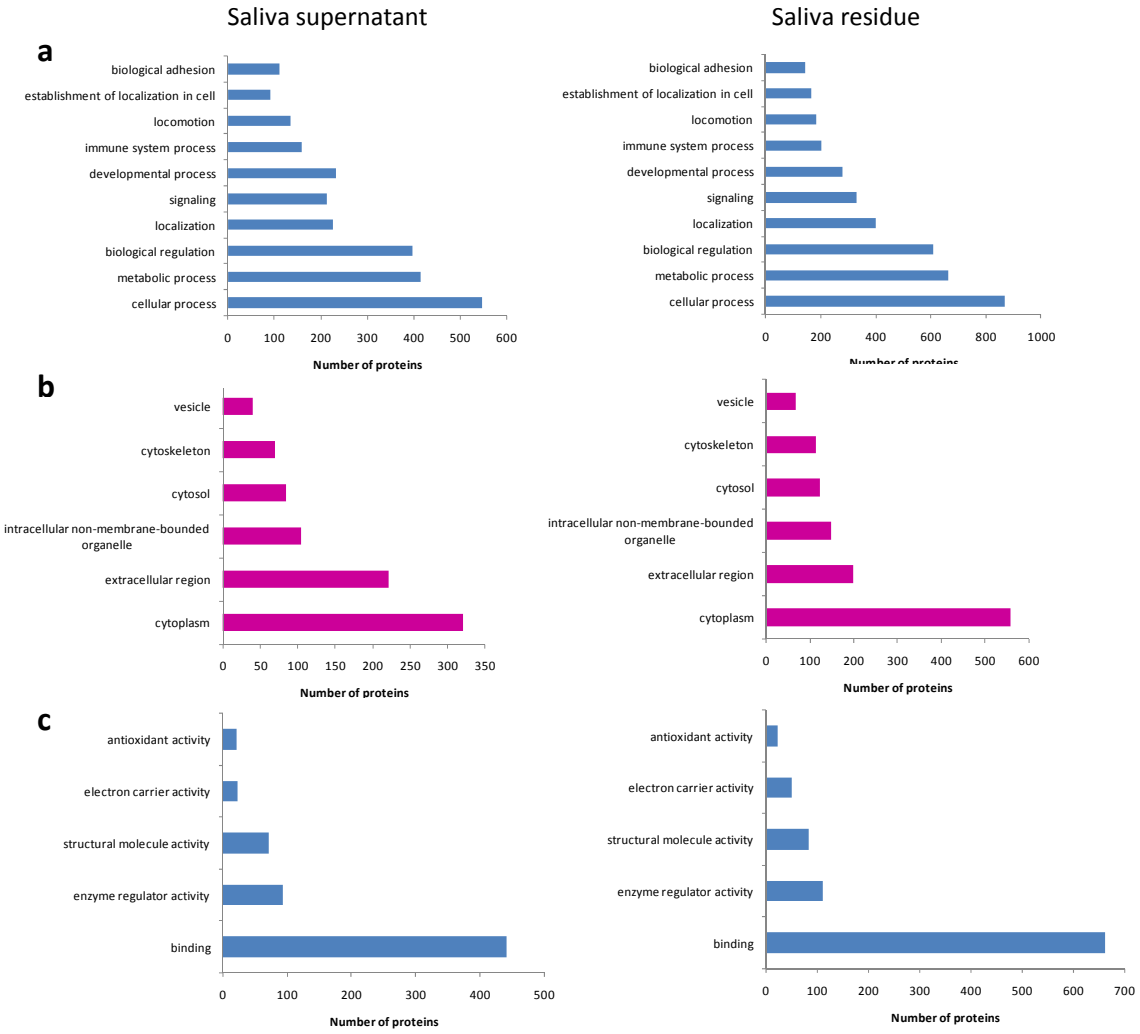


Figure 3.3 Gene Ontology annotation of human salivary supernatant (left) and residue (right) proteins. (A) GO annotation by biological process. (B) GO annotation by cellular component. (C) GO annotation by molecular function. Performed using Proteome Bioknowledge Library (Section 3.2.2).

GO enrichment analysis was performed on the identified human salivary supernatant and residue to determine which categories were significantly represented (Figure 3.4).

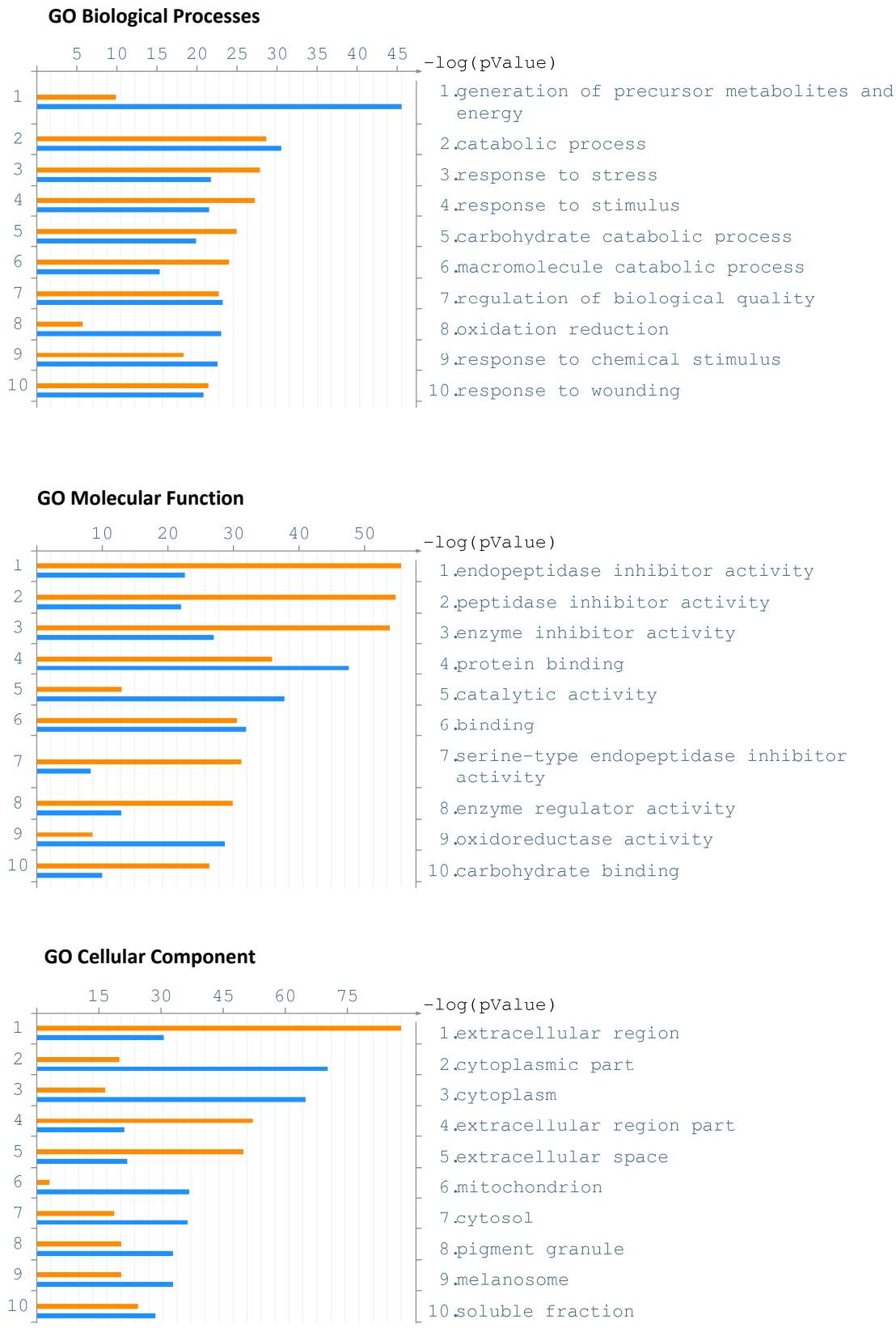


Figure 3.4 Comparison of salivary protein identification in saliva supernatant (orange) and residue (blue) using GO enrichment analysis. GO enrichment analysis was performed using MetaCore Analytical suite 4.5 (Section 3.2.2). Only the top 10 categories are displayed.

For the GO Biological Process categories, catabolic processes were the most enriched in saliva and the saliva residue. This is expected since many catabolic enzymes are present in saliva, and one of saliva's main functions is to initiate the digestion of foodstuffs. The categories 'response to wounding', 'acute inflammatory response' and 'defence response' were also highly enriched in both saliva and saliva residue. The abrasive nature of food mastication can lead to many small injuries to oral epithelial tissues and may explain the increased abundance of wound-response and inflammatory proteins. For many mammals, the licking of wounds is common practice and has been shown to speed up the healing process and reduce bacterial infection; this is consistent with the finding of saliva enriched in defense-response proteins. The categories 'response to stress' and 'response to stimulus' were also highly enriched in both the supernatant and residue. The categories 'generation of precursor metabolites and energy', 'transport' and 'establishment of localisation' were highly enriched in the saliva residue but not in saliva, presumably due to the low cellular content of saliva.

For the GO Molecular Function, the categories 'peptidase inhibitor activity' and 'enzyme inhibitor activity' were the most enriched in the saliva supernatant. 'Protein binding' was highly enriched in both the saliva supernatant and the residue. Many other 'binding' categories are highly enriched, especially in the saliva supernatant and to a slightly lesser extent in the residue. This highlights the role of salivary proteins binding to bacterial proteins and foodstuffs. 'Oxidoreductase activity' was highly enriched in the residue, whilst it was only moderately enriched in the supernatant.

For the GO Cellular Component, the category 'extracellular region' was highly enriched in the saliva supernatant, much more so than in saliva residue. The 'cytoplasm' and other cell constituents including 'mitochondrion' and 'cytosol' were more enriched in the saliva residue compared to the supernatant, as would be expected.

Since there was a large degree of overlap between the proteins observed in the saliva supernatant and residue, similar GO profiles would be expected. GO analysis was also carried out on the subset of proteins in the supernatant which were not observed in the salivary residue (Figure 3.5). The SN-only proteins were highly enriched for immune response pathways, cell adhesion, blood coagulation and proteolysis. Enrichment analysis against diseases yielded the 'pure' salivary supernatant to be enriched for proteins involved in thrombosis, amyloidosis and hypertension. The diseases for which the associated proteins were enriched are particularly interesting since they provide a starting point for potential conditions that might be diagnosed in saliva. The processes which were enriched in the SN-only proteins were much more similar to what would be expected for saliva compared to that obtained when comparing all the proteins obtained in the whole saliva supernatant. It is likely that the SN-only proteins give an indication of the proteins that are specific to the glandular secretions and do not have cellular origin.

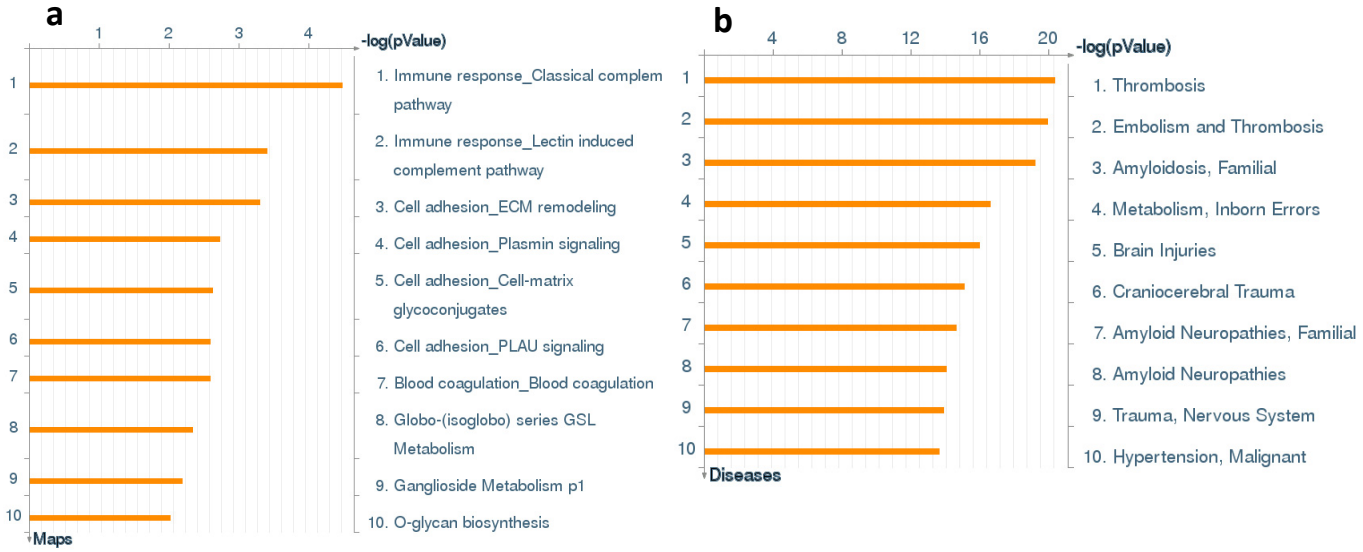


Figure 3.5 GO enrichment analysis of SN-only proteins (only those found in the salivary supernatant and not in the salivary residue). (a) Enrichment using pathway maps (GeneGo). (b) Enrichment using GeneGo diseases by biomarkers (GeneGo). GO enrichment analysis was performed using MetaCore Analytical suite 4.5 (Section 3.2.2).

3.3.1.3 Human salivary phosphoproteome

To further characterise the salivary proteome and determine whether many phosphoproteins can be identified in saliva, the phosphoprotein fraction of saliva was investigated with a view to determine whether this enrichment would be a useful step in saliva biomarker discovery. Analysis of the saliva supernatant following enrichment for phosphoproteins using IMAC identified 66 phosphorylation sites in 39 proteins (Table 3.2). This is the first report, to date, of the analysis of the salivary phosphoproteome. It is interesting to identify the phosphorylation of PRKCD, which could be an activation event, since PKC activity was observed in saliva (discussed in Section 4.3.2.2). The proteins found to be phosphorylated were investigated using enrichment analysis (Figure 3.6).

On account of the small number of proteins identified, accurate enrichment analysis was difficult. However, salivary phosphoproteins appear to be enriched for processes involving the cytoskeleton. Processes involving interleukin production and other immune processes such as neutrophil mediated immunity were also enriched. Enrichment of proteins involved in immune processes may be expected on account of the mouth (saliva) being one of the main points of entry to the body of bacterial and viral pathogens. Furthermore, the natural bacterial flora in the oral cavity must be constantly monitored and regulated by host defences, giving credence to the high abundance of immune-related proteins. Enrichment for cell adhesion is also consistent with the known interaction between salivary proteins and bacterial proteins. However, it is not immediately clear why processes involving the cytoskeleton were enriched. One possible explanation is that the cells in the mouth are highly mitotically active on account of constant cell sloughing and the turnover rate of the sulcular epithelium is one of the highest of the various epithelial tissues in the body[302].The increased cellular turnover could result in the enrichment of proteins associated with the cytoskeletal reorganisation and cell-division.

Table 3.2 Proteins in saliva with one or more phospho sites identified. Saliva was enriched for phosphopeptides using a protein peptide IMAC protocol and analysed on an LTQ Orbitrap (Section 2.2.3).

Gene Symbol	Protein name	Sites	Sequence
AMY2B	Alpha-amylase 2B	1	K.TGS#GDIENYNDATQVR.D
ARHGEF6	Rho guanine nucleotide exchange factor 6	1	R.MS#GFIYQGK.I
BTN3A3	Butyrophilin subfamily 3 member A3	1	R.ILT#LEPT#ALTICPIPK.E
DNAJC3	DnaJ homolog subfamily C member 3	1	K.LIES#AEELIR.D
EFHA1	EF-hand domain-containing family member A1	3	K.DIEDTLS#GIQTAGCGS#T#FFRDLGDK.G
FAM65B	Protein FAM65B	1	R.SQS#FAGFSGLQER.R
GREM2	Gremlin-2	1	K.EVLASS#QEALVVTER.K
HMGA1	High mobility group protein HMG-I/HMG-Y	3	R.KQPPVS#PGTALVGS#QKEPSEVPTPK.R
			R.KQPPVS#PGTALVGSQK.E
			R.KQPPVSPGT#ALVGSQK.E
HNRNPA3	Heterogeneous nuclear ribonucleoprotein A3	1	R.SSGSPY#GGGYGSGGGSGGYGSR.R
HS3ST4	Heparan sulfate glucosamine 3-O-sulfotransferase 4	1	R.TPLAPSEMITAQS#ALPER.E
HTN1	Histatin-1	4	R.EFPFYGDYGS#NYLYDN.-
			R.EFPFYGDY#GSNYLYDN.-
			R.EFPFYGDYGSNY#LYDN.-
			R.EFPFY#GDYGSNYLYDN.-
HTN3	Histatin-3	4	R.GYRS#NYLYDN.-
			R.GYRSNY#LYDN.-
			R.GYRSNY#LY#DN.-
			R.GY#RSNYLYDN.-
IQSEC3	IQ motif and SEC7 domain-containing protein 3	1	K.T#VLSVPHR.R
LCP1	Plastin-2	1	R.GS#VSDEEMMELR.E
LMO7	LIM domain only protein 7	1	R.S#WASPVYTEADGTFSR.L
LPO	Lactoperoxidase	6	R.LKTAMSS#ETPTSR.Q
			K.TAMS#SETPTSR.Q
			R.LKT#AM*SSETPTS#R.Q
			K.TAMSSETPTS#R.Q
			R.LKTAMS#SETPTS#R.Q
			R.LKTAM*S#SETPT#SR.Q
LPP	Lipoma-preferred partner	1	R.VLT#AKASTDL.-
LSP1	Lymphocyte specific protein 1	6	R.T#PSPLVLEGTIEQSS#PPLSPT#TK.L
			R.TPSPLVLEGTIEQSS#PPLS#PTTK.L
			R.TPSPLVLEGTIEQS#SPPLS#PTTK.L
			R.QAS#IELPSMAVASTK.S
MYL9	Myosin regulatory light polypeptide 9	2	R.ATS#NVFAMFDQSQIQEFK.E
			R.AT#SNVFAM*FDQSQIQEFK.E
MYO1F	Myosin-If	1	R.GPPSTSLGAS#R.R
MYO9B	Myosin-IXb	1	R.RTS#FSTSDVSK.L
NADK	NAD kinase	1	R.SLS#ASPALGSTK.E
OR2AK2	Olfactory receptor 2AK2	1	K.M*AVS#FLS#QSK.T
PIKFYVE	1-phosphatidylinositol-3-phosphate 5-kinase	1	K.ILLDS#VQLK.D
PKP1	Plakophilin-1	1	R.FSSYS#QMENWSR.H
PRAM1	PML-RARA-regulated adapter molecule 1	5	R.TSS#EPEVSVLPK.R
			R.T#SSEPEFNS#LPR.K
			R.RPS#AAS#IDLR.R
			R.RPS#AASIDLR.R
PRKCD	Protein kinase C delta type	1	R.SDS#ASSEPVGIYQGFEK.K
RAB11FIP2	Rab11 family-interacting protein 2	1	R.NNMTAS#MFDLSMK.D
ROS1	Proto-oncogene tyrosine-protein kinase ROS	2	K.YAQLLGS#WT#YTK.T
RRBP1	Ribosome-binding protein 1	1	K.LQS#SEAEVR.S
S100A9	Protein S100-A9	1	K.MHEGDEGPGHHHKPGLGEGT#P.-
SCAMP2	Secretory carrier-associated membrane protein 2	1	R.AASS#AAQGAFQGN.-
TBC1D10B	TBC1 domain family member 10B	1	R.AAGGAPS#PPPPVR.R
THRAP3	Thyroid hormone receptor-associated protein 3	1	R.IDIS#PSTFR.K
VIM	Vimentin	1	R.S#LYASSPGGVYATR.S
VTN	Vitronectin	1	R.DS#WEDIFELLFWGR.T
WIPF1	WAS/WASL-interacting protein family member 1	2	R.NLS#LSSSTPPLPS#PGR.S
			R.NLS#LSSSTPPLPSPGR.S
ZYX	Zyxin	1	R.S#PGAPGPLTLK.E

This short preliminary study of the salivary phosphoproteome indicated that several phosphoproteins are present in saliva. However, the enrichment protocol was considerably long and did not lend itself to parallel processing of samples, an important factor for biomarker discovery and analysis of clinical samples and phosphoprotein enrichment of saliva was not included in the analysis of clinical trial samples. Future in depth analysis of the salivary phosphoproteome will be performed using larger

amounts of saliva and using alternative enrichment protocols as described[303] to determine whether a greater number of phosphoproteins can be identified in saliva and whether this could be a useful fractionation technique for saliva biomarker discovery in future studies.

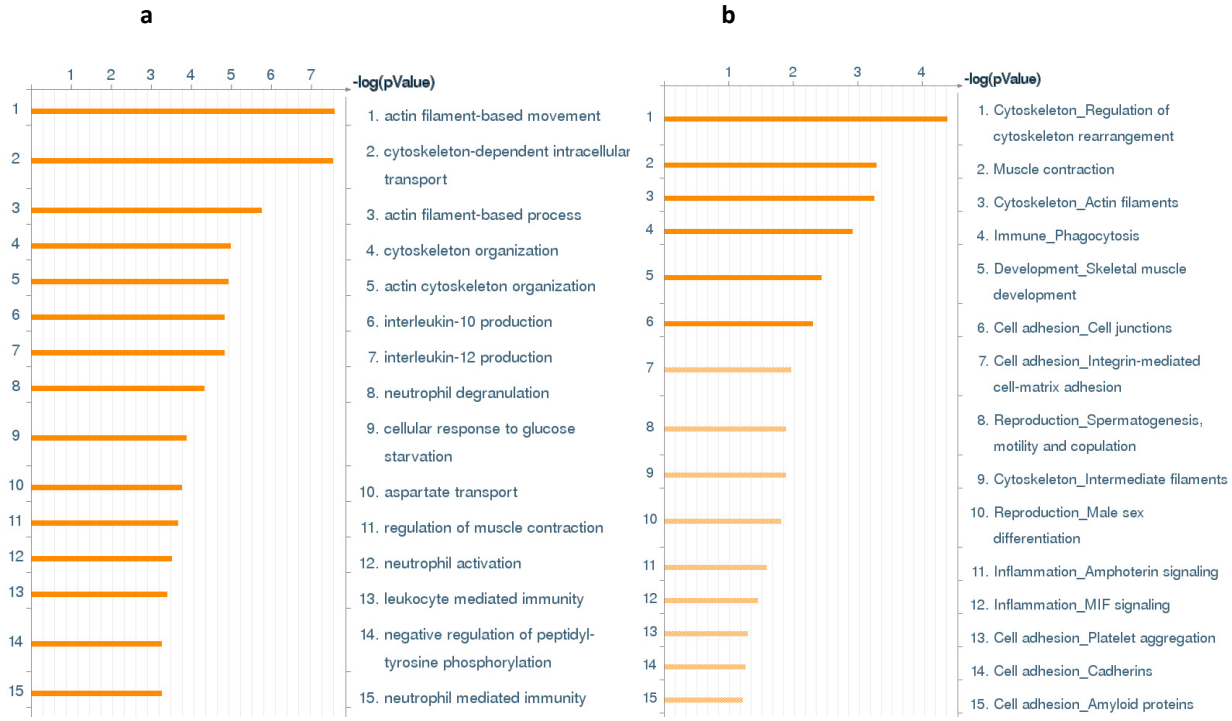


Figure 3.6 Protein annotation analysis of phosphoproteins identified in saliva supernatant. Enrichment analysis of (a) GO Biological Process and (b) GeneGo Process Networks. Analysis carried out as described in Experimental procedures.

3.3.2 Proteomics of bacteria in human saliva

Very few studies have investigated the bacterial or human cellular proteins in the saliva residue and the first proteomic study of the saliva residue was only published recently[211]. Proof-of-concept exploratory studies were therefore performed at Edinburgh to determine whether bacterial proteins could be identified in saliva using proteomics.

Various initial studies were performed in Edinburgh. From an in-solution digest of the saliva residue in one LC-MSMS analysis (HCT, Bruker) only four eubacterial proteins were identified. Obtaining a pure bacterial saliva sample was a challenge since the separation of bacterial cells from human cells and mucins was problematic. In order to increase the proportion of bacteria *vs.* human proteins to enable more detailed analysis, saliva was cultured in LB broth at 37°C for 24 hours and then analysed. This approach greatly increased the number of bacterial proteins identified and reduced the proportion of human proteins present. Following liquid culture, 87 eubacterial proteins were identified in one LC-

MSMS run (HCT, Bruker). To investigate the composition further, SCX fractionation (30 fractions) was performed on this cultured sample, resulting in 487 eubacterial proteins and 37 archaebacterial proteins. From the published proteomic literature, no archaebacteria have been reported in saliva. According to the HOMD (Human Oral Microbiome Database) one archaebacterium is reported to be present in the oral cavity[150], and no archaebacteria have been hypothesised to contribute to human disease[304]. This suggests that the bacterial flora of the mouth may be more diverse than originally anticipated, and that a proteomic analysis may provide insight into this unexplored population. Furthermore, the majority of saliva proteomics experiments have concentrated on the supernatant and have not included bacterial databases in their searches, presumably on account of the unwieldy size of the bacterial database. To the author's knowledge, no published research has been performed in which salivary proteins have been searched against archaeal databases. With a view of obtaining a 'purer' bacterial sample under more mouth-like culturing conditions, bacterial biofilm samples from an experimental mouth model (Section 1.4.1) were analysed. A total of 59 eubacterial and 6 proteins from archaea were identified in a single LC-MSMS analysis with no fractionation.

These initial proofs-of-concept studies demonstrated that saliva residue is an abundant source of bacterial proteins. A more comprehensive global analysis of the bacteria in saliva was subsequently performed using GeLC-MSMS analysis on an LTQ-Orbitrap. In the proof-of-concept work, lysis was performed using a Cell Perturbator machine (Section 2.2.2) whereas bacterial sample preparation in this global analysis involved lysis with detergents (Section 2.3.2). For the same saliva sample for which the supernatant and residue had been analysed for human proteins (Section 3.3.1.1), 20 fractions were analysed and the spectra were searched against a composite database including eubacteria, archaea and human proteins. A total of 1,365 proteins were identified with a FDR <1 % at the protein level with 6,685 unique peptides and 34,107 total peptides being identified. Of the 1,365 proteins, 1,233 matched to human, 130 matched to eubacteria and 2 matched to archaea. For the supernatant 810 proteins were identified with a FDR <1 % at the protein level with 4,001 unique peptides and 26,425 total peptides. Of the 810 proteins, 736 matched to human, 70 matched to eubacteria and 4 matched to archaea. On account of the large size of the composite database, only 1 missed cleavage was specified in the search criteria to reduce search time. This explains the slight decrease in the number of human proteins observed when the dataset was searched solely against the human proteins with a missed cleavage of 2 (1,426 proteins). Combining the bacterial identifications for residue and supernatant resulted in the proteins mapping to 97 bacterial species and 16 species with ≥ 2 proteins per species.

The majority of bacterial proteins comprising the saliva supernatant were gram-negative Proteobacteria, followed by gram-positive Actinobacteria and Firmicutes (Figure 3.7). In the saliva residue, the distribution of proteins between the different phyla was more evenly distributed. Whilst Proteobacteria also comprised the largest proportion, the contribution of both Firmicutes and gram-negative Bacteriodetes accounted for 50% of the identified proteins. Interestingly, Bacteriodetes were only observed in the residue.

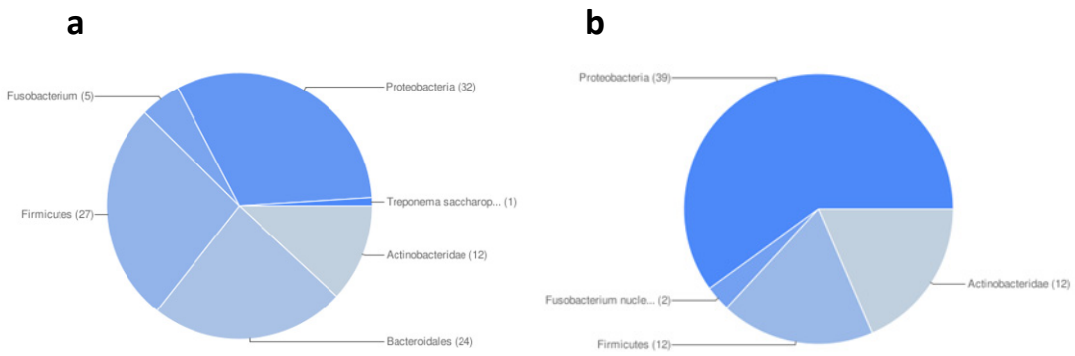


Figure 3.7 Distribution of bacterial proteins across the major phyla identified in saliva (a) residue and (b) supernatant. Taxonomic analysis performed using protein descriptions from Uniprot.

Most of the species identified would be expected to be present in human saliva[215], however, some species were unexpected. For example *Bradyrhizobium*, a soil bacterium, and *Desulfovibrio*, a water/soil bacterium were found in saliva. Since many of the organisms in the mouth have not yet had their genomes fully sequenced[215], it is likely that hits from other species may occur where the protein in different species have similar sequences. A total of 97 species were identified based on one protein and 16 species identified with at least 2 proteins for each species (Table 3.3).

This work investigating the bacterial complement of the salivary proteome indicates that many different bacterial species can be identified and that the bacterial portion of saliva may serve as a useful realm of information for future biomarker studies.

Table 3.3 Bacterial species identified in saliva residue and supernatant from one individual subject. The number of proteins and the total spectral counts for each bacterial species are given. A total of 16 bacterial species were identified with at least two proteins for each species and only those are included in the table. Spectral count refers to the number of spectra for which a peptide is successfully sequenced. The spectral count for a protein is the sum of all the spectral counts for each of the peptides identified for that protein. The number of proteins identified for a bacterial species and the total spectral count for a bacterial species can give an estimation of the abundance of the bacterial species.

	Number of proteins		Total Spectral count	
	Residue	**Supernatant**	**Residue**	**Supernatant**
Bacteroides fragilis	4	0	8	0
Bacteroides vulgatus	4	0	9	0
Fusobacterium nucleatum	4	2	8	4
Veillonella dispar	4	0	5	0
Actinomyces odontolyticus	3	6	11	12
Bacteroides caccae	3	0	13	0
Pseudomonas aeruginosa	3	1	3	2
Selenomonas flueggei	3	1	17	3
Bacillus subtilis	2	0	2	0
Bacteroides ovatus	2	0	4	0
Bacteroides uniformis	2	0	3	0
Bdellovibrio bacteriovorus	2	1	9	21
Bradyrhizobium sp.	2	1	25	1
Burkholderia xenovorans	2	1	13	4
Desulfovibrio magneticus	2	1	8	2
Porphyromonas gingivalis	2	0	6	0

3.3.3 Salivary protein complexes

Saliva was run on BN PAGE (discussed in Section 2.7.7) to identify the presence of protein complexes in saliva. In total 510 proteins were identified (0.1 % peptide FDR) from 10 GeLC-MSMS fractions. For proteins that had a spectral count greater than 5, the protein profile (spectral count) of each protein across the 10 fractions of the BN gel was compared with each of the others and the Pearson correlation coefficient was calculated (Figure 3.8 and Figure 3.9). This correlation analysis reveals proteins with similar profiles which may be present in complexes with each other. The largest molecular weight complex comprises predominately amylase and immunoglobulins (Figure 3.8d and Figure 3.10). The second largest complex comprises many enzymes (Figure 3.11). Lower molecular weight complexes were also observed (Figure 3.12 and Figure 3.13).

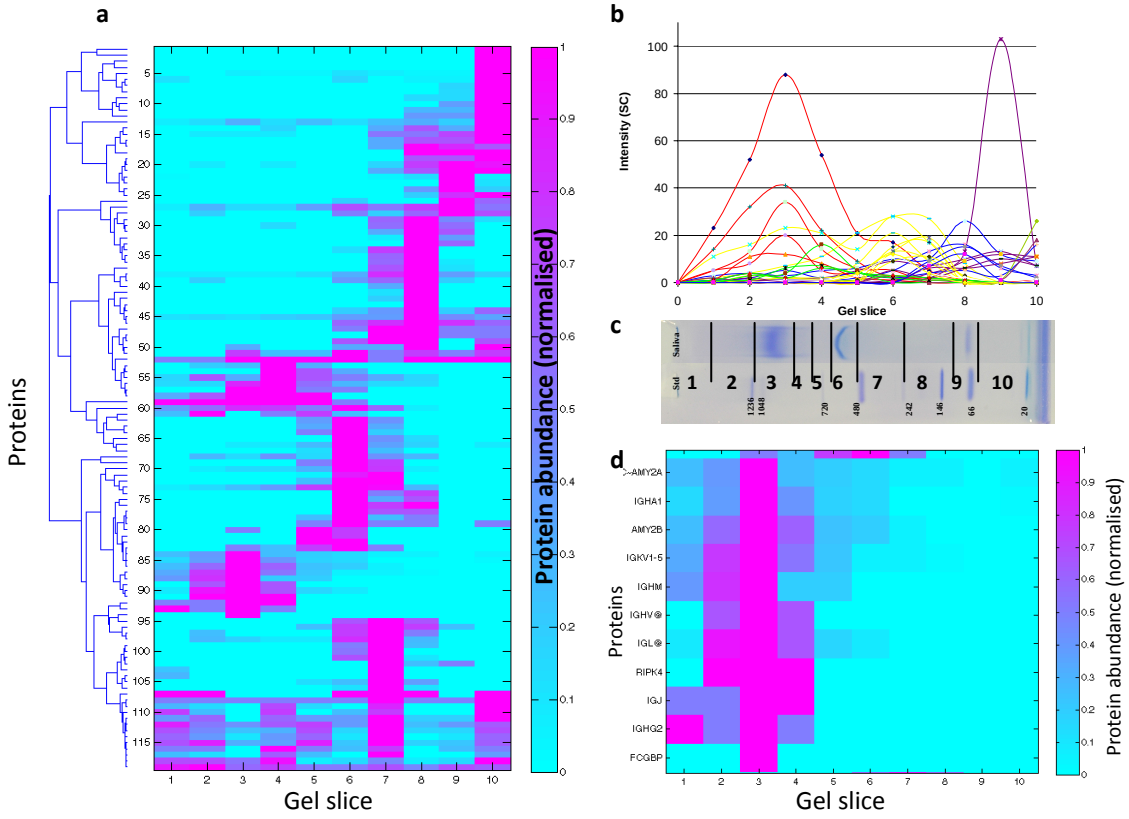


Figure 3.8 Salivary protein complexes. Saliva was run on a Blue native PAGE gel and ten slices were analysed by GeLC-MSMS. Spectral counts of the proteins give an estimate of the protein abundance in each of the fractions. (a) Profile of protein complexes for each gel slice. Hierarchical clustering grouped proteins with similar spectral count profiles together. Proteins were normalised to the gel slice with the highest number of spectral counts. Only proteins with a total spectral greater than 10 were considered. (b) Spectral count profile of individual proteins. (c) Saliva (20 µg) run on Blue native PAGE gel (Invitrogen). (d) Enlarged snapshot of (a) shows putative salivary protein complex comprising amylase and Igs with a maximum intensity in gel slice 3.

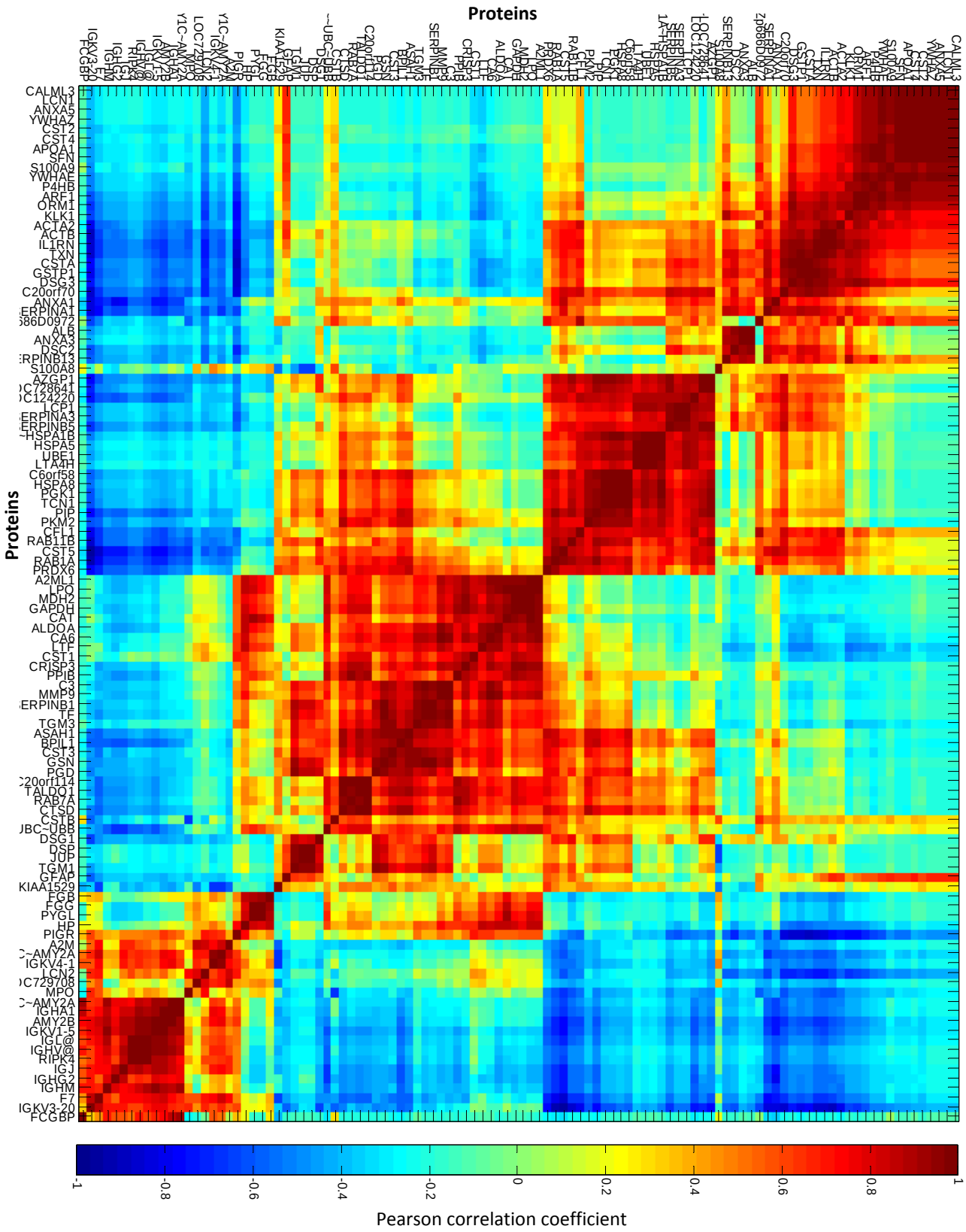


Figure 3.9 Heat map of Pearson correlation coefficients for saliva proteins analysed using BN PAGE LC-MSMS correlation profiling. Areas with high correlation coefficients (red boxes) indicate proteins that form possible complexes within human saliva.

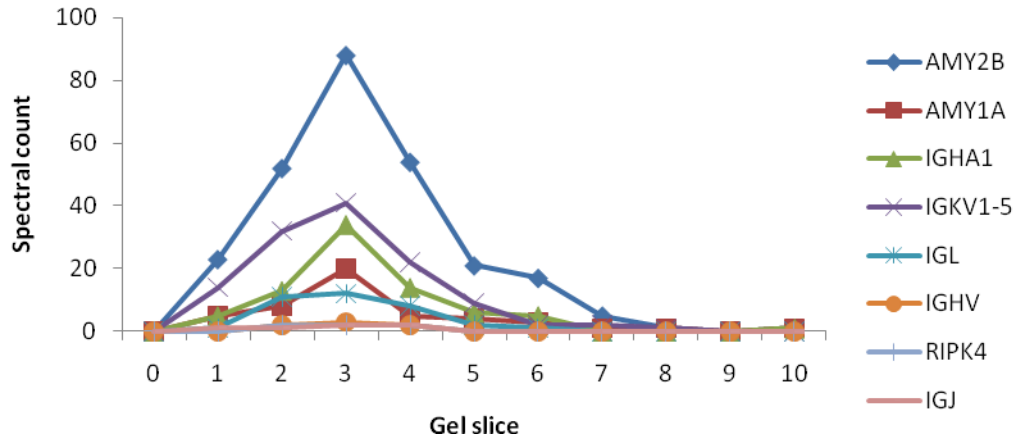


Figure 3.10 Putative salivary protein complex 1 comprising amylase, immunoglobulins and the kinase RIPK4 (receptor-interacting serine/threonine-protein kinase 4). Saliva was run on a Blue native PAGE gel and ten slices were analysed by GeLC-MSMS. Spectral counts of the proteins give an estimate of the protein abundance in each of the fractions.

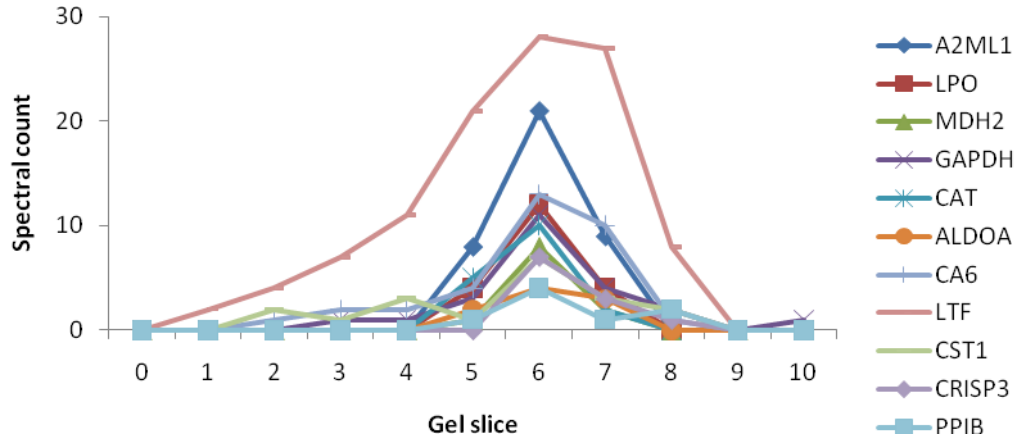


Figure 3.11 Putative salivary protein complex 2 comprising various enzymes (LPO, MDH2, GADPH, CAT, ALDOA, CA6, PP1B), A2ML1 (alpha-2-macroglobulin-like), CST1 (cystatin 1), CRISP3 (cysteine-rich secretory protein 3) and LTF (lactotransferrin). Saliva was run on a Blue native PAGE gel and ten slices were analysed by GeLC-MSMS. Spectral counts of the proteins give an estimate of the protein abundance in each of the fractions.

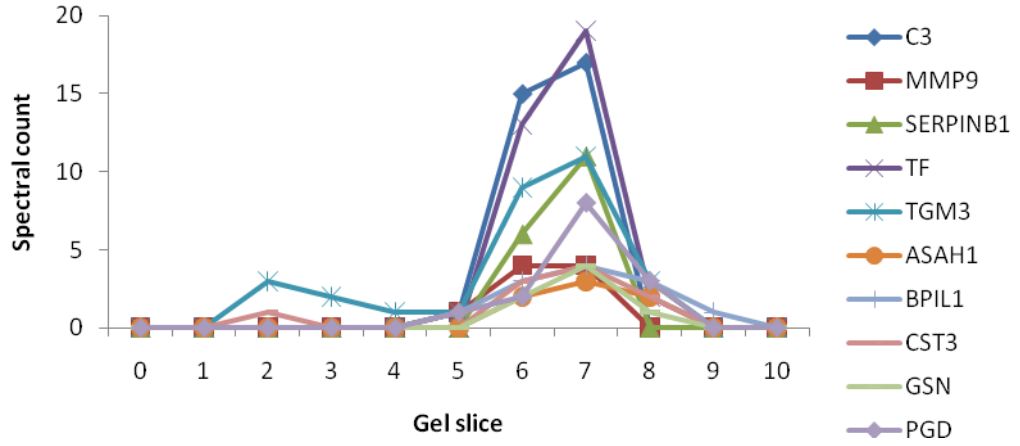


Figure 3.12 Putative salivary protein complex 3 comprising various enzymes (MMP9, TGM3, ASAH1, PGD), SERPINB1 (serpin peptidase inhibitor), CST3 (cystatin 3), C3 (complement component), TF (transferrin), BPIL1 (bactericidal/permeability-increasing protein-like 1) and GSN (gelsolin). Saliva was run on a Blue native PAGE gel and ten slices were analysed by GeLC-MSMS. Spectral counts of the proteins give an estimate of the protein abundance in each of the fractions.

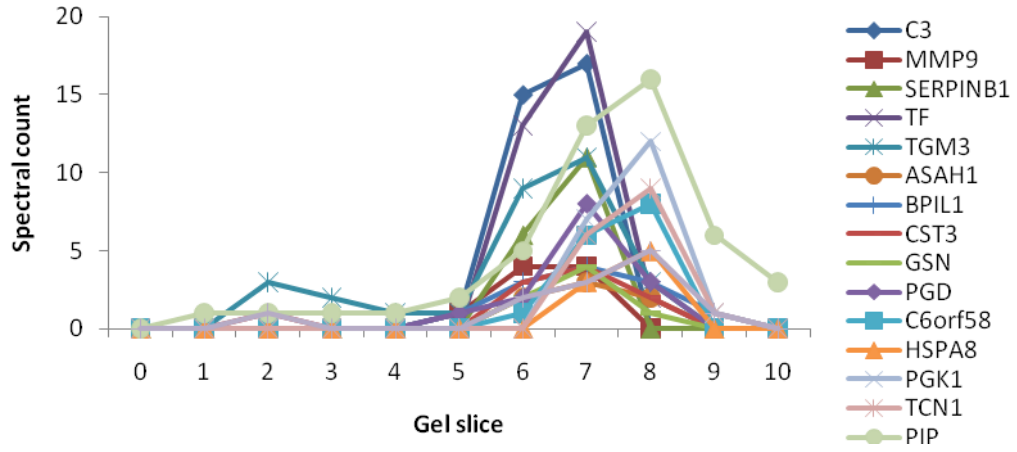


Figure 3.13 Putative salivary protein complex 4 comprising C6orf58, HSPA8, PGK1 (phosphoglycerate kinase 1), TCN1 (transcobalamin 1), PIP (prolactin-induced protein) and PKM2 (pyruvate kinase, muscle). Saliva was run on a Blue native PAGE gel and ten slices were analysed by GeLC-MSMS. Spectral counts of the proteins give an estimate of the protein abundance in each of the fractions.

For confirmation of the presence of the main protein complex in saliva (comprising amylase and immunoglobulins), a separate experiment was performed in which a different saliva sample was analysed and the gel was profiled. In this case, a complex comprising similar proteins was observed (Appendix F.1). The salivary profile from saliva run on an SDS-PAGE gel did not contain such a protein complex (Appendix F.2). To determine whether protein complexes could be observed in the saliva residue and in tear supernatant these samples were also investigated using BN PAGE protein correlation profiling and several complexes were evident (Appendix F.3).

In-gel BN PAGE kinase activity/kinase abundance correlation profiling (as described in Section 2.7.7.4) using saliva was performed but poor correlation was obtained on account of poor BN PAGE resolution and low spectral counts (Appendix G).

These experiments using correlation profiling of BN PAGE and LC-MSMS offer great potential as a technique to identify protein complexes. Whilst 2D BN/SDS PAGE has been widely used as a method to investigate protein-protein interactions, very little work using 1D BN PAGE in combination with proteomics has been performed[275, 305-306] and only one recent publication reported the use of BN PAGE linked with protein correlation profiling to identify protein complexes in mitochondria[306]. Whilst the small precast BN PAGE used in this study may not be suitable for large scale mapping of protein complexes, there is potential for greater resolution and loading capacity with larger BN gels.

Whilst BN PAGE protein correlation profiling was shown to be useful for the identification of protein complexes and could be useful for small biomarker studies in the future comparing a few samples, the procedure requires sample fractionation and therefore would currently be too time consuming for high throughput biomarker discovery with a large number of samples.

3.3.4 Comparison of label-free quantitation software packages

With a view to develop a biomarker platform, three label-free software packages were evaluated: MS-Xelerator, Progenesis and XCMS. Prior to using these softwares to compare saliva samples, less complex mixtures were used.

3.3.4.1 Experimental design to compare three label-free quantitation software packages

Three different datasets were used to assess the packages with varying levels of complexity: standard peptides (3 different concentrations), BSA digest, and saliva spiked with BSA. Two different saliva samples were used to rigorously test the software packages. Significant differences in the saliva protein composition can be observed between wake-up (WU) and after-toothbrushing (T). These differences were observed on SDS-PAGE (Figure 3.14), where T saliva had a more intense band around 75 kDa. The goal was to optimise the assay using samples of different complexity to then apply it on healthy versus gingivitis saliva from clinical trials.

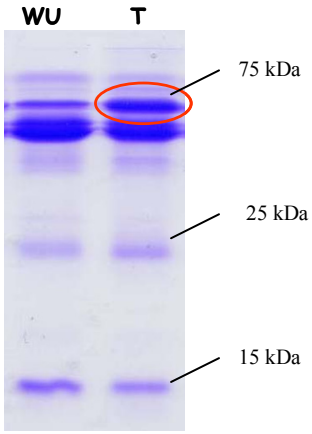


Figure 3.14 SDS-PAGE of wake-up (WU) and after-toothbrushing (T) saliva samples. Saliva supernatant (20 µg) was loaded. A more intense band at 75 kDa in the T saliva was present.

The aim was to perform analysis of all the samples (peptides, BSA, saliva spiked with BSA, saliva WU+T) by four methodological approaches (20 and 150 min gradients on a low and high resolution mass spectrometer) using four different quantitation approaches (manual, MSX, Progenesis, XCMS) (Table 3.4). By varying the complexity of the datasets, accurate assessment of the dynamic range of each software package can be determined. By analysing simple datasets, where the correct fold values were known, the parameters were optimised and subsequently applied to the more complex datasets where the fold values were not known. Due to time constraints, not all of these combinations were obtained before moving from Edinburgh to HMS. Those datasets shown with an 'X' were collected (Table 3.4). Manual analysis was very time consuming and was therefore only performed for the low complexity samples.

Table 3.4 The different samples, instrument set-ups and data analysis combinations that were performed for a comparison of 3 label-free software packages: MS-Xelerator (MSX), Progenesis (Prog) and XCMS. An 'X' refers to the datasets that were collected and analysed. NA refers to datasets that were too complicated for manual analysis. WU refers to saliva that was taken at wake-up and T refers to saliva taken following tooth brushing.

MS	HCT								FTICR					
Gradient length	20 min				150 mins				20 min			150 min		
Data analysis	Manual	MSX	Prog	XCMS	Manual	MSX	Prog	XCMS	MSX	Prog	XCMS	MSX	Prog	XCMS
Peptides	X	X	X	X										
BSA digest	X	X	X	X	X	X	X	X	X	X	X			
Saliva spiked with peptides	X	X	X	X										
Saliva spiked with BSA	X	X	X	X	X	X	X	X	X	X	X	X		X
Saliva (WU vs T)					NA	X	X	X				X	X	X

There are a variety of settings and processing options for the 3 label-free software packages (Table 3.5). In general, MSX and Progenesis were fairly similar in terms of ease of use and both had a good graphical interface. Progenesis had a significant advantage in that many groups could be compared, whilst only two groups could be compared in MSX and XCMS. At the time of analysis, linking of LCMS quantitation data with LC-MSMS was only possible with the MPeaks module of MSX and with Progenesis. Unfortunately, the MPeaks module only allowed the comparison of two LC-MSMS runs and not two groups of samples as in the MSCompare module. Integration of the LCMS and LC-MSMS linking is currently being implemented into the MSCompare module. The earlier version of Progenesis, for which only a beta version was available in the laboratory, did not have LC-MSMS linking capability. However, the later version allowed this linking capability through importing the Mascot results file.

Table 3.5 Summary of the three label-free quantitation software packages. The software packages MSX, Progenesis and XCMS software packages were evaluated.

	MSX	**Progenesis**	**XCMS**
File import form	NetCDF/mzXML	NetCDF/mzXML	NetCDF/mzXML
Operating system	Windows	Windows	Linux, OSX, Windows
Type	commercial	Commercial	Open source (R)
Normalisation	TIC area, specific peaks, specified value	Based on unchanging peptide peaks	None
LCMS and LC-MSMS linking capability	no	yes	no
reference	www.MsMetrix.com	www.nonlinear.com	http://metlin.scripps.edu/download/
Preprocessing	Despiking	Smoothing/Baseline correction automatic	Smoothing (Gaussian) – can change the fwhm
	Smoothing (Gaussian/Savitzky-Golay		
	Baseline correction		
	Deisotoping not possible	Deisotoping possible	Deisotoping not possible
Alignment	Offset, correlation optimized warping (COW), reference peak warping (RPW) – combination of above possible	Vector based – automatic and manual adjustment possible	Diffferent algorithms (peak detection): - binlin - bin - intline - cwave
			- Can change the 'missing value' which determines how many points used for alignment.
Group settings	Can only compare 2 groups	Can compare many groups	Can only compare 2 groups
Biomarker searching	Can change sensitivity (S/N)		S/N can be changed
	Possible to change the range over which biomarkers are searched	Not possible to change the range over which biomarkers are searched	Not possible to change the range over which biomarkers are searched
	Can change the 'resolution' and fshm of the biomarker map – determines time taken to generate map		
Output data format	Exportable to excel tables		
	Biomarker map and alignment images can be saved.	All images can be saved as .png and report generated automatically	png graphics of EIC and alignment can be generated.

3.3.4.2 Standard peptides

Peptides at three different concentrations (150, 300 and 600 fmol) were analysed on a low resolution ion trap mass spectrometer (HCT, Bruker) with a gradient of 20 min. A mixture of 5 peptides comprising bradykinin, met-enkephalin, angiotensin II, glu-fibrinopeptide B and melittin was used (Figure 3.15) and each concentration was analysed in triplicate LCMS runs.

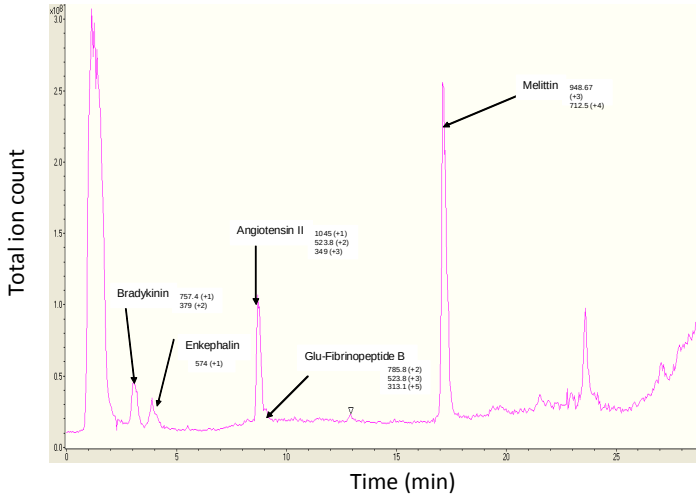


Figure 3.15 Total ion chromatogram of LCMS run with standard peptides (HCT, 20 minute gradient). The peaks for the 5 standard peptides are shown, with the charge stages which were observed in the MS.

Analysis of the average fold of individual peptide peaks was determined using the three software packages as well as manually using the peak area of the extracted ion intensity (Table 3.6). The results of the manual analysis and using Progenesis were very similar, with averages very close to the theoretical 2.0 and small standard deviations. A much greater number of peaks were identified using MSX and XCMS analysis since the software packages do not apply de-isotoping. The melittin peptide was not identified using Progenesis, therefore only 4 peaks were identified. The peak-picking algorithms of MSX and XCMS were much more sensitive; however a disadvantage of the ability to pick up very low abundance peaks is reduced accuracy and analysis of peaks which may not represent peptides and may be from noise. If only the peaks for the largest charge stage of each of the 5 peptides were chosen, then the accuracy increased (resulting in an average fold value closer to 2) and smaller standard deviations were obtained (Table 3.7). Only the peak area values can be obtained using Progenesis and XCMS whilst for MSX, both the peak height and peak area can be obtained. One of the advantages of Progenesis is that the area of the selected peak can be manually edited, although clearly for large numbers of samples this is not practical. A p-value was generated by both Progenesis and XCMS. However, MSX at the time of analysis did not provide any statistical measure of error. Some of the peaks detected using XCMS were outliers and skewed the results, despite having P values of <0.01. If these peaks were excluded (following manual checking of the EIC alignment), the accuracy of the average fold value increased. Whilst this manual checking is possible for a small number of peaks, it is not feasible for global analysis.

Table 3.6 Comparison of average fold values for standard peptides analysed on HCT with a 20 min gradient. Triplicate LCMS analyses with 3 different concentrations were analysed manually and with the 3 different software packages acquired using the HCT ion trap MS.

[peptides] (fmol)	**Manual**			**MSX**			**Progenesis (P<0.01)**			**XCMS (P<0.01)**			**XCMS (P<0.01) some peaks excluded following manual inspection**		
	Peaks	Av	SD	Peaks	Av	SD	Peaks	Av	SD	Peaks	Av	SD	Peaks	Av	SD
300/150	5	1.94	0.11	57	2.26	0.45	4	2.02	0.28	29	2.50	1.64	27	2.14	0.54
600/300	5	1.93	0.05	52	2.13	0.39	4	1.97	0.11	42	3.78	5.69	35	2.11	0.39

Table 3.7 Comparison of individual peptide fold values for standard peptides acquired using the HCT with a 20 min gradient. Triplicate LCMS analyses with 3 different concentrations were analysed manually and with the 3 different software packages acquired using the HCT ion trap MS.

		Charge state	**Manual**		**MSX**		**Progenesis**	**XCMS**
Peptide	m/z		Average fold (peak area)	Average fold (peak height)	Average fold (peak area)	Average fold (peak height)	Average fold (peak area)	Average fold (peak area)
Bradykinin	379	2+	1.84	2.01	2.01	1.97	1.90	1.95
Enkephalin	574	1+	2.05	1.79	2.11	1.84	1.89	2.00
Angiotensin 2	524	2+	1.85	1.68	2.11	1.98	1.85	1.92
Melittin	712	4+	1.91	1.78	2.21	2.00	NF	1.84
GFP	786	2+	2.07	1.85	3.27	2.14	2.44	NF
Average			1.94	1.82	2.34	1.99	2.02	1.93
SD			0.11	0.12	0.52	0.11	0.28	0.07

3.3.4.3 BSA digest

As an increase in complexity from a 5 peptide mixture, a BSA digest at 2 different concentrations, was analysed using both the HCT and FTICR mass spectrometers. The fold change of individual BSA peptide peaks was determined using each of the software packages and the 9 most intense BSA digest peaks were used for manual analysis (Figure 3.16).

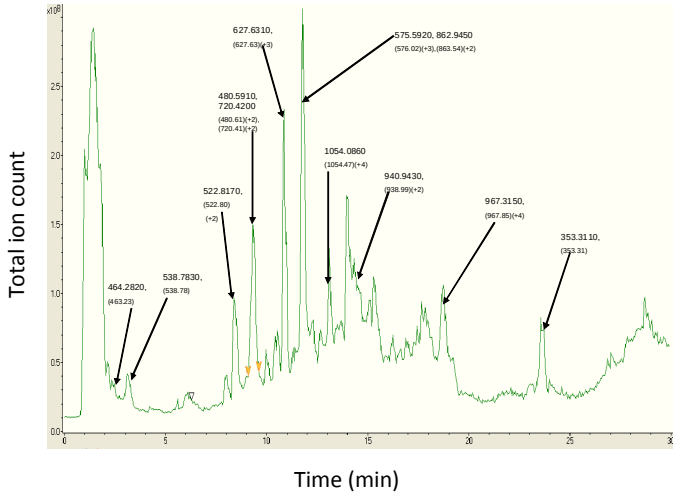


Figure 3.16 Total ion chromatogram of LCMS run with BSA digest (HCT, 20 minute gradient). The peaks for the most intense BSA digest peptides (black arrows) were used for manual quantitation.

Similar to the five standard peptide experiment, very few peaks (only nine) were detected using Progenesis. Conversely, 123 and 269 peaks were detected by XCMS and MSX respectively (Table 3.8). Analysis of the BSA digest using Progenesis was not as successful as the analysis of the five peptide sample and the fold value was underestimated, and the standard deviation was large. Many more peaks were identified using both MSX and XCMS and both had similar fold values and standard deviations. Nine intense BSA peptide peaks were chosen for manual validation and these were compared in the three software packages (Table 3.9). In some cases, the software packages did not always detect the nine peptides. However, when only these peptides were considered, all the software packages performed well with very similar averages and standard deviations obtained for the detected peptides.

Table 3.8 Comparison of average fold values for BSA digest analysed on HCT with a 20 min gradient. Triplicate LCMS analyses with 3 different concentrations were analysed manually and with the 3 different software packages acquired using the HCT ion trap MS.

[BSA] fmol	**Manual**			**MSX**			**Progenesis (P<0.01)**			**XCMS (P<0.01)**			**XCMS (P<0.01) some peaks excluded following manual inspection**		
	Peaks	Av	SD	Peaks	Av	SD	Peaks	Av	SD	Peaks	Av	SD	Peaks	Av	SD
200/100	9	1.97	0.29	269	2.18	0.41	9	1.35	0.74	123	3.03	9.25	122	2.12	0.45

Table 3.9 Comparison of individual peptide fold values for BSA digest using the HCT with a 20 min gradient. Triplicate LCMS analyses with 3 different concentrations were analysed manually and with the 3 different software packages acquired using the HCT ion trap MS.

BSA peptides Rt (min)	**m/z**	**Charge state**	**Manual**		**MSX**		**Progenesis**	**XCMS**
			Average fold (peak area)	Average fold (peak height)	Average fold (peak area)	Average fold (peak height)	Average fold (peak area)	Average fold (peak area)
8.4	523	2+	1.86	1.72	1.93	1.94	2.09	1.97
9.4	481	2+	2.02	1.89	1.87	1.71	NF	NF
9.4	720	2+	2.33	1.97	2.23	2.10	2.12	2.31
11.0	628	3+	2.21	1.64	NF	NF	NF	NF
11.7	576	3+	1.98	2.03	2.01	1.95	NF	2.25
11.8	864	2+	2.06	2.02	2.21	2.16	2.03	2.20
13.1	1054	4+	1.96	1.95	1.78	1.87	1.82	1.93
14.1	939	2+	1.29	1.18	2.18	1.56	NF	NF
18.7	968	4+	2.02	2.25	2.12	2.19	NF	NF
Average			1.97	1.85	2.04	1.94	2.02	2.13
SD			0.29	0.31	0.17	0.22	0.14	0.17

The same BSA digest samples were also analysed on the high resolution FTICR mass spectrometer (Table 3.10). Strikingly, the number of peaks identified by MSX and XCMS was far lower whilst a greater number of peaks were identified by Progenesis, compared to the analysis with the low resolution HCT. For this sample on the high resolution FTICR, the analysis using XCMS was the most optimal with an average close to 2 and a small standard deviation. Very large standard deviations

were observed for both MSX and Progenesis. When examining individual peptides, the MSX peak height measurements were inaccurate on account of poor alignment of the spectra. On examining the individual BSA peptides (Table 3.11) many of the peptides were not observed, although the standard deviations were smaller than those when considering all the peaks detected. The lower number of peaks identified by MSX and XCMS was likely to be due to the lower sensitivity of the FTICR compared to the HCT. Using a higher concentration of BSA, a greater number of peaks were detected by all three software packages (Table 3.12). Whilst analysis using MSX and XCMS resulted in similar averages, around 2, the analysis using Progenesis resulted in a very large inaccurate fold of 12 and a very large standard deviation on account of poor chromatographic alignment.

Table 3.10 Comparison of average fold values for BSA digest analysed on FTICR with a 20 min gradient. Triplicate LCMS analyses with 3 different concentrations were analysed manually and with the 3 different software packages acquired using the FTICR MS.

[BSA] fmol	**MSX**			**Progenesis (P<0.01)**			**XCMS (P<0.01)**		
	Peaks	Av	SD	Peaks	Av	SD	Peaks	Av	SD
200/100	48	2.84	1.56	16	2.32	2.87	11	1.98	0.35

Table 3.11 Comparison of individual peptide fold values for BSA digest using the FTICR with a 20 min gradient. Triplicate LCMS analyses with 3 different concentrations were analysed with the 3 different software packages acquired using the FTICR MS.

BSA peptides Rt (min)	**m/z**	**Charge state**	**MSX**		**Progenesis**	**XCMS**
			Average fold (peak height)	Average fold (peak height)	Average fold (peak area)	Average fold (peak area)
8.4	523	2+	2.38	3.02	2.17	1.73
9.4	481	2+	2.47	4.31	2.97	NF
9.4	720	2+	1.76	1.79	1.96	1.75
11.0	628	3+	NF	NF	NF	NF
11.7	576	3+	NF	NF	2.32	1.90
11.8	864	2+	2.50	2.54	NF	NF
13.1	1054	4+	NF	NF	NF	NF
14.1	939	2+	NF	NF	NF	NF
18.7	968	4+	NF	NF	NF	NF
Average			2.28	2.92	2.36	1.79
SD			0.35	1.06	0.44	0.09

Table 3.12 Comparison of average fold values for BSA digest at 1 and 2 pmol on FTICR with a 20 min gradient. Triplicate LCMS analyses with 3 different concentrations were analysed with the 3 different software packages acquired using the FTICR MS.

[BSA] pmol	**MSX**			**Progenesis (P<0.01)**			**XCMS (P<0.01)**		
	Peaks	Av	SD	Peaks	Av	SD	Peaks	Av	SD
2/1	191	2.10	0.63	518	12.4	44.9	364	1.94	0.39

Table 3.13 Comparison of individual peptide fold values for BSA digest at 1 and 2 pmol using the HCT with a 20 min gradient. Triplicate LCMS analyses with 3 different concentrations were analysed with the 3 different software packages acquired using the HCT ion trap MS.

BSA peptides Rt (min)	m/z	Charge state	MSX		Progenesis	XCMS
			Average fold (peak area)	Average fold (peak height)	Average fold (peak area)	Average fold (peak area)
8.4	523	2+	2.24	2.24	1.79	1.91
9.4	481	2+	2.08	2.08	2.07	1.90
9.4	720	2+	1.62	1.62	1.88	1.72
11.0	628	3+	NF	NF	2.80	1.86
11.7	576	3+	NF	NF	NF	NF
11.8	864	2+	NF	1.51	1.64	1.43
13.1	1054	4+	NF	1.72	3.63	1.50
14.1	939	2+	NF	NF	20.6	1.95
18.7	968	4+	1.91	1.91	6.26	1.71
Average			1.96	1.85	5.08	1.75
SD			0.27	0.28	6.45	0.20

The experiments using high concentrations of BSA were in the pmol range (Table 3.13). However, as a result of the relatively high concentrations, the trap in the FTICR mass spectrometer was overloaded. For some of the peaks, particularly the most intense ones, the intensity did not increase proportionally with the concentration and a plateau was observed (Figure 3.17), accounting for the lower fold values often observed.

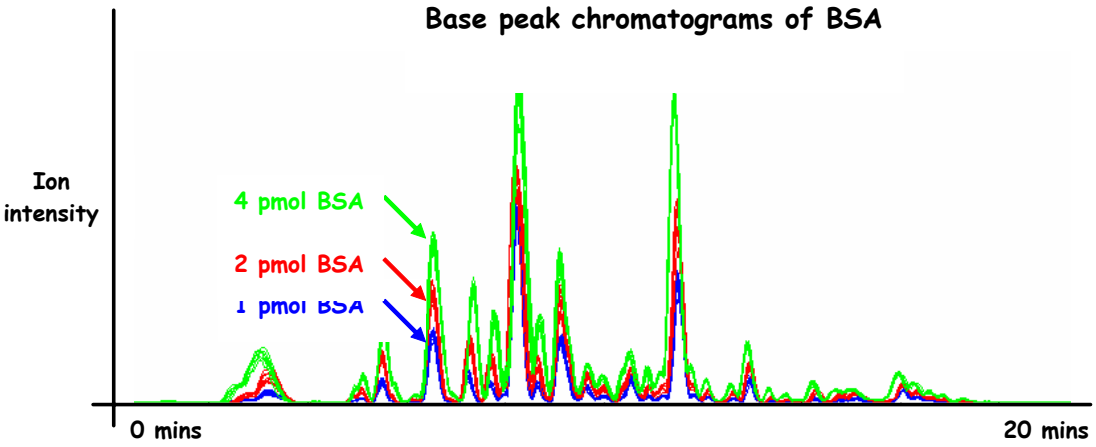


Figure 3.17 Base peak chromatograms of three different concentrations of BSA digest. Alignment of the individual LCMS runs was performed using MSX. Different concentrations of BSA digest (1, 2 and 4 pmol) were analysed by LCMS using a 20 min gradient on a FTICR MS. Using the high concentrations of BSA, for some of the peptides, the ion trap is saturated and the ion intensity does not increase linearly with the increased concentration of BSA digest.

3.3.4.4 Saliva spiked with peptides and BSA

For analysis of biological samples, it is important to be able to detect changes in a few proteins amongst many unchanging compounds. To model this scenario, digested saliva (a complex mixture) was spiked with two different concentrations of standard peptides and a BSA digest (Figure 3.18).

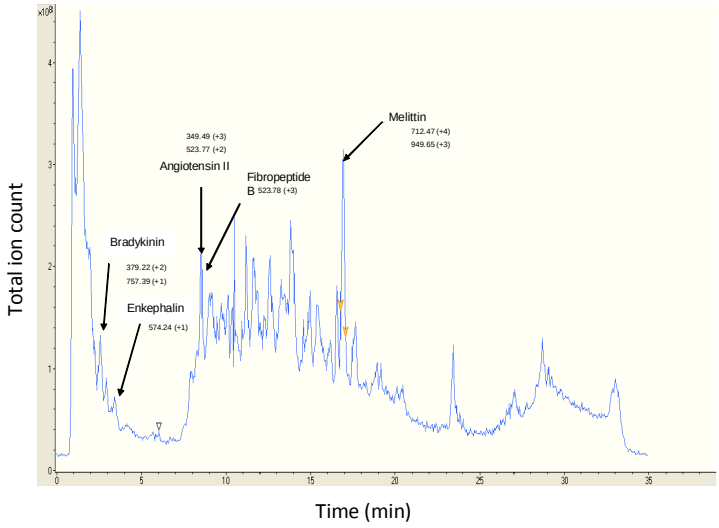


Figure 3.18 Total ion chromatogram of LCMS run with digested saliva spiked with a standard peptide mix (bradykinin, met-enkephalin, angiotensin II, glu-fibrinopeptide B and melittin). LCMS was performed in the HCT with a 20 minute gradient. The peaks for the spiked 5 standard peptides are indicated with black arrows.

Focusing solely on the 5 standard peptides, the averages and standard deviations of the three software packages were fairly similar to the analysis 'by hand' (Table 3.14). The peptides were spiked in at 150 and 300 fmol resulting in a theoretical average of two. The fold values observed were slightly less than two, presumably on account of ionisation suppression with the presence of saliva peptides. Melittin and bradykinin were not identified using Progenesis. The process where peaks are picked using Progenesis is based on the observation of a charge state distribution. In the case of the 4+ melittin charge state there may have been difficulty in observing this due to the low resolution data from the HCT. However, it is not clear why the bradykinin peak was not detected.

Table 3.14 Comparison of individual peptide fold values for saliva spiked with standard peptides using the HCT with a 20 min gradient. Triplicate LCMS analyses with 3 different concentrations were analysed manually and with the 3 different software packages acquired using the HCT ion trap MS.

Peptide	m/z	**Charge state**	**Manual**		**MSX**		**Progenesis**	**XCMS**
			Average fold (peak area)	Average fold (peak height)	Average fold (peak area)	Average fold (peak height)	Average fold (peak area)	Average fold (peak area)
Bradykinin	379	2+	1.70	1.32	1.84	1.87	NF	2.02
Enkephalin	574	1+	1.75	1.55	1.71	1.72	1.70	1.66
Angiotensin 2	524	2+	1.85	1.69	1.65	1.67	1.67	1.63
Melittin	712	4+	1.44	1.78	1.62	1.42	NF	1.58
GFP	786	2+	1.65	1.77	1.79	1.68	1.96	1.60
Average			1.68	1.62	1.72	1.67	1.78	1.70
SD			0.15	0.19	0.09	0.16	0.16	0.18

For saliva spiked with two different concentrations of BSA, for the 9 BSA digest peaks, many peaks were not identified by Progenesis (Table 3.15). Smoothing was not used on the peaks and the added complexity of the saliva background resulted in lower abundance peaks which were more challenging to quantitate. In contrast, analysis using MSX and XCMS, which both use preprocessing smoothing algorithms, had smaller standard deviations. However, when the sample was run on a longer gradient, 150 min rather than 20 min, the accuracy of manual analysis improved considerably (Table 3.16). The analysis using Progenesis did not improve with the longer gradient and the quality of quantitation using XCMS decreased due to poorer alignment with the longer gradient.

Table 3.15 Comparison of individual peptide fold values for saliva spiked with BSA digest using the HCT with a 20 min gradient. Triplicate LCMS analyses with 3 different concentrations were analysed manually and with the 3 different software packages acquired using the HCT ion trap MS.

BSA peptides Rt (min)	m/z	Charge state	Manual		MSX		Progenesis	XCMS
			Average fold (peak area)	Average fold (peak height)	Average fold (peak area)	Average fold (peak height)	Average fold (peak area)	Average fold (peak area)
8.4	523	2+	1.95	1.71	2.31	1.95	1.64	1.93
9.4	481	2+	1.78	1.58	1.56	1.76	NF	1.95
9.4	720	2+	1.84	1.62	1.83	2.06	NF	1.73
11.0	628	3+	0.88	0.90	NF	NF	NF	1.74
11.7	576	3+	0.92	1.12	1.84	1.93	NF	NF
11.8	864	2+	0.75	0.92	1.91	2.10	1.61	1.51
13.1	1054	4+	NF	NF	2.24	2.04	NF	1.56
14.1	939	2+	NF	NF	1.90	2.03	NF	1.85
18.7	968	4+	1.64	1.5	1.18	1.64	NF	1.75
Average			1.39	1.34	1.85	1.92	1.63	1.75
SD			0.52	0.35	0.36	0.19	0.02	0.16

Table 3.16 Comparison of individual peptide fold values for BSA digest using the HCT with a 150 min gradient. Triplicate LCMS analyses with 3 different concentrations were analysed manually and with the 3 different software packages acquired using the HCT ion trap MS.

m/z	Charge state	Manual		MSX		Progenesis	XCMS
		Average fold (peak area)	Average fold (peak height)	Average fold (peak area)	Average fold (peak height)	Average fold (peak area)	Average fold (peak area)
523	2+	1.81	1.54	NF	NF	NF	1.71
481	2+	1.78	1.41	1.63	1.40	NF	1.59
720	2+	1.88	1.22	1.44	1.33	NF	1.47
628	3+	NF	NF	1.96	1.80	NF	NF
576	3+	1.45	1.61	2.09	1.78	NF	NF
864	2+	2.08	1.83	2.04	1.76	1.96	NF
1054	4+	1.88	1.54	1.88	1.74	NF	NF
939	2+	NF	NF	1.79	1.68	NF	NF
968	4+	1.97	1.71	1.82	1.91	NF	NF
Average		1.84	1.55	1.83	1.68	1.96	1.59
SD		0.20	0.20	0.22	0.20		0.12

Using the FTICR to analyse the 9 BSA peptide peaks, Progenesis analysis identified only three peptides whilst MSX and XCMS identified five (Table 3.17).

Table 3.17 Comparison of individual peptide fold values for BSA digest using the FTICR with a 20 min gradient. Triplicate LCMS analyses with 3 different concentrations were analysed with the 3 different software packages acquired using the FTICR ion trap MS.

BSA peptides Rt (min)	m/z	Charge state	MSX	Progenesis	XCMS
			Average fold (peak height)	Average fold (peak area)	Average fold (peak area)
8.4	523	2+	NF	NF	NF
9.4	481	2+	2.33	1.95	2.13
9.4	720	2+	1.84	1.6	1.87
11.0	628	3+	NF	NF	NF
11.7	576	3+	NF	NF	NF
11.8	864	2+	2.35	1.90	1.88
13.1	1054	4+	1.95	NF	2.04
14.1	939	2+	1.16	NF	1.09
18.7	968	4+	NF	NF	NF
Average			1.93	1.82	1.80
SD			0.48	0.19	0.41

3.3.4.5 Two saliva samples compared: 'wake-up' versus 'after-toothbrushing'

As the final level of 'complexity', two different saliva samples were compared: 'wake-up' (WU) and 'after-toothbrushing' (T). For both the HCT and FTICR, MSX detected the greatest number of peaks (Table 3.18 and Table 3.19). For Progenesis, about 16 times as many peaks were detected using the FTICR as with the HCT. With the higher resolution, the isotopic distribution was better resolved and the charge states were assigned more easily. With the lower resolution HCT, where the isotopic distribution could not be resolved, the peak was detected as 'noise' since a charge state was not defined. For XCMS, the reverse was observed and the number of peaks detected was around 60 times more for the HCT than for the FTICR. In all the cases, the number of peaks with a fold change greater two between WU and T was high and very close to the number of peaks detected (Figure 3.19). Manual inspection of the chromatograms showed poor alignment of the spectra for all the software packages, accounting for the large number of differential peaks observed.

Table 3.18 Comparison of two saliva digest samples acquired using the HCT with a 150 min gradient. LCMS analyses of saliva digest samples taken at wake-up (WU) and after toothbrushing (T) were analysed using using the 3 different software packages. Duplicate LCMS runs were acquired.

	MSX	Progenesis (P<0.01)	XCMS (P<0.01)
Number of peaks	1385	25	1040
Peaks> 2 fold change between WU and T	1356	24	896

Table 3.19 Comparison of two saliva digest samples acquired using the FTICR with a 150 min gradient. LCMS analyses of saliva digest samples taken at wake-up (WU) and after toothbrushing (T) were analysed using using the 3 different software packages. Duplicate LCMS runs were acquired.

	MSX	Progenesis (P<0.01)	XCMS (P<0.01)
Number of peaks	719	410	16
Peaks> 2 fold change between WU and T	579	389	16

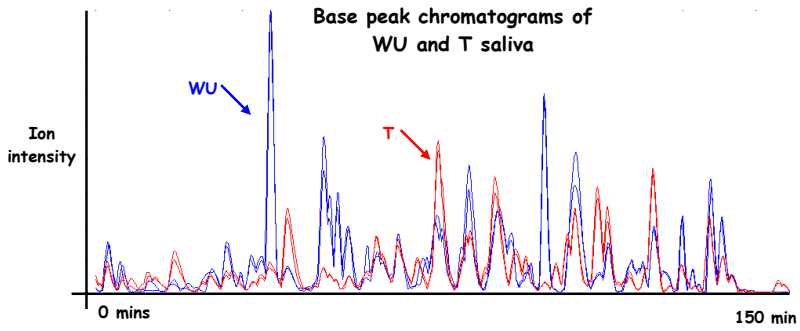


Figure 3.19 Base peaks chromatograms of WU and T saliva. LCMS analyses were acquired using FTICR MS and with a 150 min gradient and duplicates of each saliva sample were performed. Alignment of the base peak chromatograms was performed using MSX.

3.3.4.6 Overall evaluation of software packages

Initially, analysis using each software package was performed using the default parameters, which resulted in poor analytical results, predominately due to poor alignment leading to incorrect ratios. In order to improve peak detection and alignment, the parameters for each software package were optimised using simple samples for which the absolute ratios were known. A variety of samples of different complexities were analysed on low and high resolution mass spectrometers with short and long gradients, and processed with the 3 software packages as well as manually (for the simple samples). Following optimisation (for more details see Appendix H) a final set of parameters was determined (Table 3.20 to Table 3.23).

Table 3.20 MSX parameters used for HCT for 20 and 150 min LC gradients. A list of the preprocessing, alignment and biomarker map variables investigated and the optimised parameters used is given.

Dataset	Parameters investigated		Optimised parameters used	Exceptions
20 min	Preprocessing	Despiking: 0-5 Smoothing: 0-10 scans Baseline corrections: 0-15	Despiking: 0 Smoothing: 4 scans Baseline corrections: 0	
	Alignment	RPW: EIC/BPC COW: (5-30)/(30-100)	RPW 0.049 interpolating segment/0.3 min max shift using BPC followed by COW: 30/5	
	Biomarker map variables	Sensitivity: 0.1-5% Minimum peak width: 0.1-1 min Peak area/height for quantitation	Sensitivity: 1% Minimum peak width: 0.3 min	For simple samples (peptides/BSA) min. peak width of 0.3 min used. For complex samples 0.1 min used.
150 min	Preprocessing	Despiking: 0-5 Smoothing: 0-10 scans Baseline corrections: 0-15	Despiking: 0 Smoothing: 20 scans Baseline corrections: 0	
	Alignment	RPW: EIC/BPC COW: (5-30)/(30-100)	RPW 0.049 interpolating segment/1 min max shift using BPC followed by COW: 50/10	For W vs T COW used for alignment: 100/15. For Saliva/BSA 50/10 COW used.
	Biomarker map variables	Sensitivity: 0.1-5% Minimum peak width: 0.6-2.6 min Peak area/height for quantitation	Sensitivity: 1% Minimum peak width: 0.8 min	Minimum peak width required very depemdent on particular dataset. 0.6 or 0.8 mins used.

Table 3.21 MSX parameters used for FTICR for 20 and 150 min LC gradients. A list of the preprocessing, alignment and biomarker map variables investigated and the optimised parameters used is given.

Dataset	Parameters investigated		Optimised parameters used	exceptions
20 min	Preprocessing	Despiking: 0-5 Smoothing: 0-10 scans Baseline corrections: 0-15	Despiking: 0 Smoothing: 2 scans Baseline corrections: 0	
	Alignment	RPW: EIC/BPC COW: (5-30)/(30-100)	RPW 0.049 interpolating segment/0.3 min max shift using BPC followed by COW: 50/10	
	Biomarker map variables	Sensitivity: 0.1-5% Minimum peak width: 0.1-1 min Peak area/height for quantitation	Sensitivity: 1% Minimum peak width: 0.1 min	For simple samples (peptides/BSA) min. peak width of 0.3 min used. For complex samples 0.1 min used.
150 min	Preprocessing	Despiking: 0-5 Smoothing: 0-10 scans Baseline corrections: 0-15	Despiking: 0 Smoothing: 5 scans Baseline corrections: 0	
	Alignment	RPW: EIC/BPC COW: (5-30)/(30-100)	RPW 0.049 interpolating segment/1 min max shift using BPC followed by COW: 50/10	Fot W/T two COWs used for alignment 100/10 and then 50/10.
	Biomarker map variables	Sensitivity: 0.1-5% Minimum peak width: 0.6-2.6 min Peak area/height for quantitation	Sensitivity: 1% Minimum peak width: 0.8 min	Minimum peak width required very dependent on particular dataset. 0.6 or 0.8 mins used.

Table 3.22 XCMS parameters used for HCT for 20 and 150 min LC gradients. A list of the user input variables investigated and the optimised parameters used is given.

Dataset	Parameters investigated		Optimised parameters used
20 min	Alignment	Bin, binlin, intlin, cwave Missing value 0-10	Bin Missing value = 1
	Gaussian / fwhm	10-80	Fwhm = 40 seconds
150 min	Alignment	Bin, binlin, intlin, cwave Missing value 0-10	Bin Missing value = 1
	Gaussian / fwhm	60-280	Fwhm = 160 seconds

Table 3.23 XCMS parameters used for FTICR for 20 and 150 min LC gradients. A list of the user input variables investigated and the optimised parameters used is given.

Dataset	Parameters investigated		Optimised parameters used
20 min	Alignment	Bin, binlin, intlin, cwave Missing value 0-10	Bin Missing value = 1
	Gaussian / fwhm	10-80	Fwhm = 40 seconds
150 min	Alignment	Bin, binlin, intlin, cwave Missing value 0-10	Bin Missing value = 1
	Gaussian / fwhm	60-280	Fwhm = 220 seconds

3.3.4.6.1 Analysis using low resolution LCMS data

For comparison of different concentrations of simple standard peptide mixtures and digested BSA analysis using all three software packages was fairly good with detected peptide fold changes near the theoretical. However, several outlier peaks which were very poorly aligned were often detected with XCMS and this adversely affected the results. For simple mixtures, manual validation removed poorly aligned peaks. However, for more complex samples, manual validation is not feasible. With peptide and BSA spiked saliva, many of the spiked in peptides and BSA peptides were not detected using Progenesis for both the short and long gradient. The analysis using MSX and XCMS was better with

these set of spiked in samples. Similarly, for the comparison of the two saliva samples, only 25 peaks were detected using Progenesis whereas MSX and XCMS detected around 1,000. For simple samples on a low resolution instrument, analysis using Progenesis appears to be the most accurate but suffers from poor sensitivity. MSX or XCMS would be the better choice for more complex samples where peak detection is of greater importance than accuracy.

3.3.4.6.2 Analysis using high resolution LMS data

For the comparison of BSA digest samples, at the lower (fmol) concentrations, the best results were obtained using XCMS in terms of accuracy, but results were poor in terms of sensitivity (very few peaks were detected). For the higher concentration (pmol) of BSA, analysis using XCMS yielded the best results of the three software packages and the sensitivity was also improved (364 peaks detected). The increase in performance for XCMS with the data obtained from the FTICR data was particularly significant. This improvement was not as significant for MSX. For Progenesis it is not strictly correct to compare the data analysis from the low and high resolution runs since different Progenesis algorithms were used for both the low resolution and high resolution data. Focusing on specific peaks, Progenesis can perform well, but the alignment of peaks across multiple datasets is very challenging.

3.3.4.6.3 General comments

Analysis using Progenesis relied on detecting an isotopic peak distribution for peak picking. Consequently, for low-resolution instruments, it was difficult to determine the isotopic distribution for charge states greater than 3+, and hence the number of peaks observed was lower than the number identified using MSX and XCMS. For example, for the comparison of WU and T saliva with analysis using Progenesis, around 400 peaks were detected on the FTICR, compared to only 25 on the HCT. Conversely, neither XCMS nor MSX rely on the isotopic distribution to determine whether a peak is 'real' or 'noise'. XCMS integration was performed on spectra that have had no background subtraction, which accounts for the detection of more peaks compared to Progenesis in which a background subtraction was performed. In MSX, there were various options for preprocessing and background subtraction. The use of profile, rather than centroided data, was optimal for XCMS and Progenesis whilst there was no significant difference between using profile or centroided data for analysis using MSX. The analysis using XCMS was more optimised for high resolution FTICR data than low resolution ion trap data. For analysis using both Progenesis and MSX, due to the very visual nature of the software, it was easy to quickly manually validate the alignment for each of the EIC peaks in the biomarker tables generated. For peaks that were misaligned or where noise had inadvertently been picked up, it was easy to exclude such peaks. Despite the ability to generate the EIC for each peak, XCMS is much less user friendly, and manually validating every peak and correcting the table is very time consuming. This manual element of checking is very useful and the visual interface of both MSX and MS-Xelerator allow this to be achieved easily. Progenesis takes this a step further and allows the user to modify the area over which the peak is integrated whereas manual adjustment is not possible with MSX or XCMS. MSX is much more flexible and can therefore be easily adapted to different data types and lengths of gradients. Unlike Progenesis and MSX, XCMS

currently does not have any normalisation functionality. This was particularly an issue where spray reproducibility is poor and the TIC varied significantly. Despite the disadvantages of XCMS, the most accurate alignment was obtained using FTICR MS coupled with analysis by XCMS, and this combination was ultimately decided to be optimal.

3.3.4.7 Workflow for biomarker discovery: LCMS and LC-MSMS linking

Having determined empirically that the FTICR mass spectrometer coupled with data analysis using XCMS performed best for label-free quantitation, the question was how to use this combination most effectively in biomarker discovery. In order to identify protein differences, it was necessary to combine the quantitative LCMS data with the corresponding LC-MSMS. The FTICR had the advantages of higher resolution and hence improved chromatogram alignment for label-free quantitation. However, the disadvantage of the FTICR is that the duty cycle is long and consequently not suited to LC-MSMS runs. To overcome this limitation, the high mass accuracy and resolution of the FTICR was combined with the fast MSMS scan rate of the HCT ion trap. The same LC system was used to separate the same sample on both the FTICR and the HCT (Figure 3.20). The LC system was on a trolley and was 'wheeled' between the two instruments as appropriate.

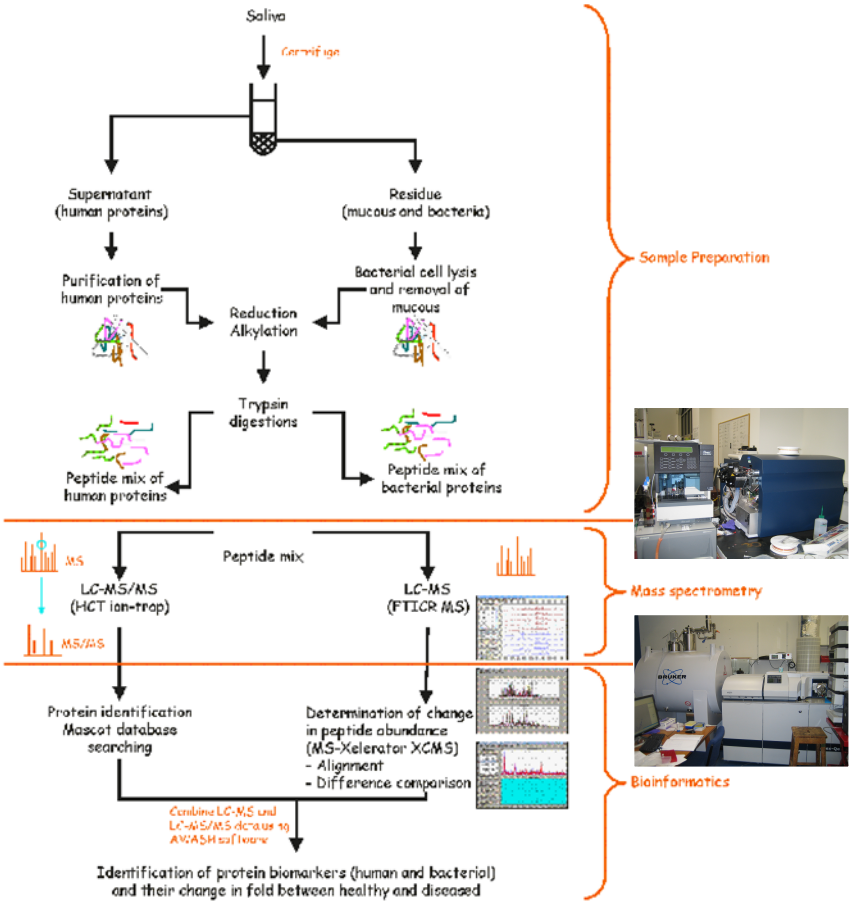


Figure 3.20 Possible workflow for biomarker discovery. Saliva supernatant and residue are digested in solution and the digests are analysed using both the FTICR and HCT mass spectrometers. LCMS quantitation is performed using FTICR whilst identification is performed using LC-MSMS on the HCT.

3.3.4.8 Combining LCMS and LC-MSMS data

Various approaches were assessed in combining LCMS and LC-MSMS data. As previously mentioned, there was no LCMS-LC-MSMS linking capability built into XCMS at that time, and the linking capability was still not fully developed for MSX and Progenesis. An in-house solution was therefore sought. A script, known as 'AWASH', written by Dr Holger Husi in the Barran lab, was used to match peaks from the aligned LCMS data (Figure 3.21) with the LC-MSMS peaks. The AWASH algorithm (the binary is obtainable from the Proteomic Analysis DataBase PADB[307]) used iterative searches of m/z and retention time pairs and matched the peaks from the Mascot LC-MSMS output with the corresponding peaks from the LCMS spectra (output from XCMS). Briefly, a starting window of 0.1 Da and 30 s was used and this window was then increased with each round of iteration for unmatched datapoints. A maximum of 1 Da for the m/z window and 120s for the retention time window was set. For XCMS data pairs that matched to several peptides in the Mascot results file within a set window, the best match was returned. However, major challenges with the alignment were experienced and a great deal of effort was put into trying to improve the reliability of the chromatographic separation, including using the Advion Replay ProtoMate.

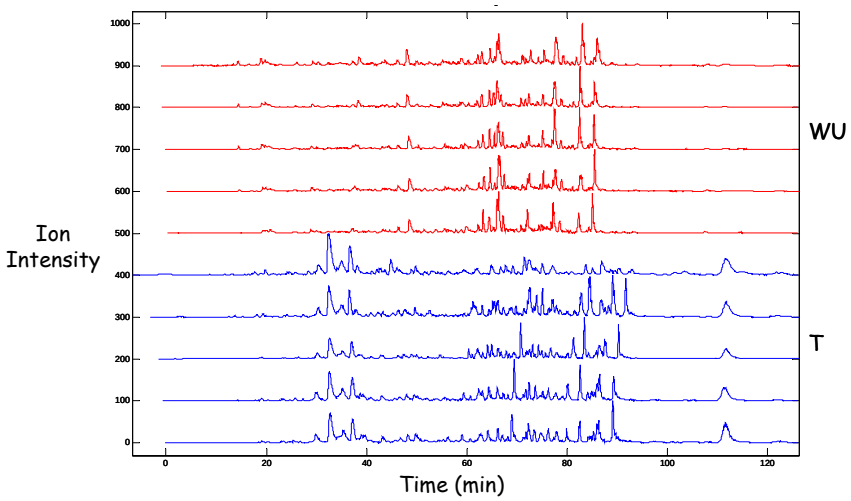


Figure 3.21 Alignment of base peak chromatograms of WU and T saliva. Five LCMS runs of each sample following alignment using MSX. Whilst the high intensity peaks can be aligned, alignment of the lower intensity peaks was not very accurate.

3.3.4.9 Advion Replay ProtoMate

Because two chromatographic runs are never exactly the same, accurate alignment of chromatogram traces is critical for accurate label-free quantitation. There were major issues with trying to align the spectra, hence the original reason for assessing several software packages. To try to address this issue, an Advion ProtoMate Replay system was assessed (Figure 3.22). The ProtoMate Replay enabled the collection of two identical nanoLC separations from a single injection. This was achieved by means of a post-column split and a capture loop.

Despite the use of a refocusing nanoLC column, the resulting chromatographic peak width increased. Another challenge was the technical difficulty in adjusting the chromatographic peak in order to optimise timing. Yet another disadvantage with the set-up using the Proxeon's Easy nanoLC and HCT/FT combination was that a series of runs could not be queued due to the direct injection column setup (without a precolumn). Since there was no pre-column, successive samples could not be loaded onto the column while the replay was in operation, otherwise flow through the capture loop would have been disrupted. As a result, the replay could not be permanently kept in the 'replay' position. Using the replay with a precolumn may have avoided the majority of these problems. A more reproducible chromatographic separation was achieved by using two different injections and, since these runs could be set up to run automatically, this was preferred over the replay approach.

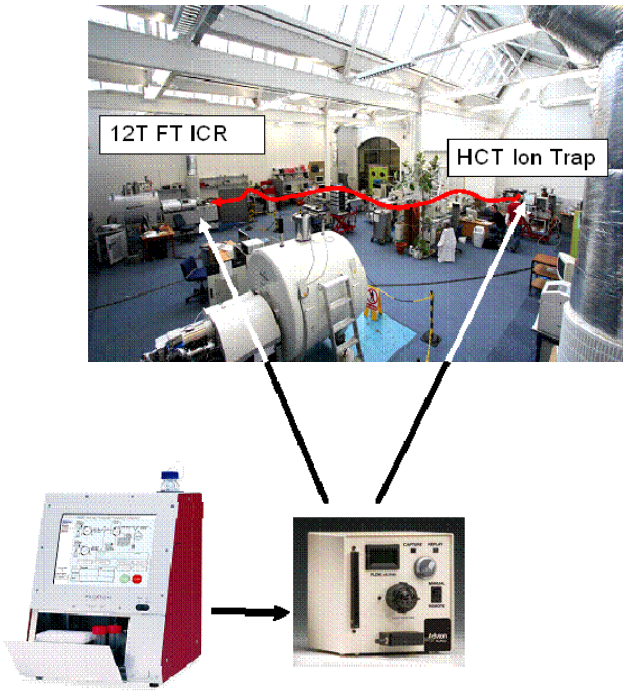


Figure 3.22 Replay set-up with HCT and FTICR. An 11 metre storage loop was connected between the HCT and the FTICR. Photo courtesy of Dr. Mackay.

For accurate alignment it was critical to have stable retention times across the replicate LCMS analyses. Many of the inaccuracies and challenges experienced were through inaccurate alignment of the chromatographic traces. This can partly be blamed on instrumental set up as well as on inadequate alignment algorithms. Ultimately, despite a large portion of time dedicated to optimising LC conditions and software parameters, the label-free approach using extracted ion chromatograms was not used for biomarker discovery and quantitation based on spectral counting was used instead.

3.4 Conclusions

The aim of the work in this chapter was to develop a platform for salivary biomarker discovery prior to the analysis of clinical trial saliva samples. Various fractions of saliva were investigated including the supernatant, residue and the phosphoprotein portion of the supernatant for both human and bacterial proteins. The presence of protein complexes in saliva was also investigated and several large complexes were identified. Three different software packages were evaluated to determine which was the most appropriate for the data analysis.

This is the first global proteomics study performed on the saliva residue and supernatant from the same test subject. The large overlap between the proteins observed in the supernatant and the residue indicate that many of the proteins in saliva originate from lysed cells from the inside of the mouth, and that this proportion is more significant than previously thought. This is also the first report of a human salivary phosphoproteome. Whilst this exploratory experiment confirmed the presence of a variety of phosphoproteins in saliva, the phosphoprotein enriched fraction of saliva will not be used for the clinical trial samples on account of the time consuming sample preparation steps. However, with the finding of a large number of bacterial proteins in saliva, both the human and bacterial proteins will be investigated further and both the saliva supernatant and the residue will be analysed for the clinical trial samples.

Furthermore, the global analysis of salivary protein complexes using proteomics presented here has not previously been performed. This technique provides a good platform which can be applied to other systems for the analysis of protein complexes in other body fluids or tissues. In this report, only complexes comprising human proteins were considered. Interactions between salivary proteins and bacteria are also critical and this system could serve as a useful platform to extend investigation to such interactions and complexes. However, one of the problems with such an approach is that only the very strong interactions between the salivary proteins and bacteria are retained after sample centrifugation, washing and cell lysis. Excluding the wash step of the pellet could help to retain such interactions

After determining that both the saliva supernatant and residue fractions will be used for biomarker discovery and that both the human and bacterial proteins will be considered, evaluation of possible software packages for quantitation was performed.

Three label-free software packages were examined: MSX, Progenesis, and XCMS using several proteomic datasets with different degrees of complexity (peptides and BSA digest at different concentrations, saliva spiked with standard peptides and BSA and two different saliva samples) on low and high resolution mass spectrometers. XCMS and Progenesis gave much better results for high resolution data acquired on the FTICR versus data from the ion trap (HCT), whilst there was no significant difference between low and high resolution data processed using MSX. Whilst all the software packages can process simple samples successfully and can achieve correct fold values for

specific peaks, the challenge was in the optimisation of the parameters for accurate alignment and to achieve a cut-off between significant peaks and noise. For complex samples it was useful to check the alignment of the extracted ion chromatograms manually and ensure that the peaks were not noise. The visual and user-friendly nature of MSX and Progenesis made this much easier than in XCMS. All the software packages were beta versions and were still undergoing development, particularly in the algorithms for chromatographic alignment and LCMS and MSMS linking capability.

For accuracy and best alignment, XCMS combined with data from the high resolution FTICR provided the most accurate workflow for label-free quantitation. However, the long duty time of the FTICR made LC-MSMS impractical. Attempts were made to try to combine the LCMS spectra of the FTICR with the LC-MSMS of the HCT, but due to the difficult nature of this task, no acceptable alignment was obtained. In practice and with the state of development of the software at the time, the envisioned workflow combining high resolution quantitative FTICR LCMS data with low resolution MSMS protein identification was not successful. As a result of the extensive challenges experienced with chromatographic alignment, label-free quantitation using the ion intensity was abandoned and a spectral counting approach was used instead to compare the saliva samples from the clinical trials. The results of the comparison of healthy and gingivitis saliva are discussed in Chapter 5.

4 Kinase ActivitY Assay for Kinome Profiling

4.1 Introduction

A strategy for a multiplexed *in vitro* kinase activity assay was recently described by the Gygi laboratory[150-151]. This Chapter outlines subsequent development of the original kinase assay and its applications to a range of tissue types and organisms. The original Kinase ActivitY Assay for Kinome analysis (KAYAK) assay[150] used 90 substrate peptides (K90) covering a set of core signalling pathways and sites identified from large scale phosphoproteomic studies[176, 308-309]. One disadvantage of the K90 set was peptide redundancy as many of the peptides were phosphorylated by the same kinase. Several hundred different peptides were synthesised, purified and evaluated, and from these a core set of 60 peptides (K60) was selected. The K60 comprises a subset of the original K90 peptides and also includes a number of new peptides that expand the range of pathways covered. The new K60 set was evaluated in 12 different human cell lines and various tissue types. To determine the general applicability of KAYAK in organisms other than humans, KAYAK was performed in *Drosophila*, mouse, and yeast. As an extension of the salivary work presented earlier, KAYAK was performed on saliva to ascertain whether kinase activity could be observed in this body fluid. The application of KAYAK to biomarker discovery in saliva will be further discussed in Chapter 5.

As gene and protein expression are very different between tissues in the same organism, a comprehensive proteomic and phosphoproteomic study was performed to characterise protein expression in nine mouse tissues[310]. Large differences in protein expression and phosphorylation were observed between the different tissues. To complement this work, kinase activity profiling was performed on mouse tissue lysates and compared to the kinase abundance in the different tissues.

Several kinases may be responsible for the phosphorylation of any one substrate peptide used in the KAYAK assay, depending on the sample. A standard biochemical technique to determine the 'active component' of a mixture is to separate that mixture chromatographically and to analyse each of the fractions for activity. Using a similar strategy it is possible to identify the kinase responsible for phosphorylation of a particular peptide. The lysate is fractionated by column chromatography and a profile of the kinase activities for each fraction is determined using KAYAK. The same fractions are also analysed by shotgun proteomics and a profile of protein abundance for each fraction is determined. Through correlating the kinase activity profiles with those of the kinase abundance, the responsible kinase can be determined. A kinase activity/kinase abundance correlation profiling experiment using the K60 substrate set was performed with HeLa nocodazole arrested cell lysate and the kinase activity profiles were correlated with the kinase abundance profiles. In addition to profiling the K60 kinase activities, selected peptides from a large scale phosphoproteomic study which were hyperphosphorylated during mitosis (G2/M) were investigated using kinase activity/kinase abundance profiling to determine putative kinases that could be responsible for their phosphorylation.

4.2 Experimental procedures

4.2.1 Kinase activity assay for kinome profiling

The general protocols for the multiplexed kinase activity assays are described in Experimental Procedures Section 2.7.3. Individual experimental conditions and deviations from the general KAYAK procedure are given below for each of the experiments discussed in this chapter.

4.2.1.1 Kinase activity profiling on 12 human cell lysates

KAYAK was performed in triplicate on 20 µg of lysate from 12 human cell lines. The classical eppendorf format KAYAK reaction was used. HeLa cells were grown and lysed as described in Section 2.7.2. Three other cell lysates (U-87 MG (glioma), HT-1080 (fibrosarcoma) and DU 145 (prostate cancer)) were obtained from Dr. Kubota, a post doctoral fellow in the Gygi laboratory, and prepared in the same fashion. Other cell lysates were obtained from slightly different growing conditions: MCF7 (breast cancer) grown in DMEM media 10% FBS and 0.01mg/ml bovine insulin, T-47D (breast cancer) cells from RPMI-1640 medium with 10% FBS and 0.2 µg/ml bovine insulin. PC-3 (prostate cancer) cells were maintained in F-12K medium with 10% FBS. Jurkat (human T lymphocyte) cells were maintained in RPMI1640 medium with 10% FBS. SUM-159 (breast cancer) cells were maintained in Ham's F-12 medium with 10% FBS, 5 µg/ml bovine insulin and 1 µg/ml hydrocortisone.

4.2.1.2 Kinase activity profiling on HEK 293T cells with overexpression of kinases

Kinase activity profiling was performed on lysates with overexpressed kinases to determine substrate specificity. Of the 22 kinases overexpressed using Gateway cloning (Section 2.6.2.2) only 11 were efficiently transfected: AKT1, MAPK14, MAPKAPK3, PAK2, PKN1, PLK1, PRKCB1, PRKCI, PRKCZ, PBK and CDC2. Kinase profiling using K60 substrates was performed in triplicate on 20 µg of each of the lysates with overexpressed kinases.

4.2.1.3 Kinase activity profiling on various organisms

Kinase activity profiling was performed on lysates from various organisms to determine whether KAYAK could be used in organisms other than humans, KAYAK was performed on 20 µg of human cell lysates (Section 2.7.2), yeast lysate, mouse tissue homogenates and *Drosophila* embryo lysates (Appendix I) .

4.2.1.4 Kinase activity profiling on saliva

Kinase activity profiling was performed on saliva to determine whether kinase activity was present in saliva and the variability between individuals. KAYAK was carried out on 10 µg of lysate from the saliva supernatant and residue from 6 human subjects. The 96 well format KAYAK reaction was used.

4.2.1.5 Kinase activity profiling on 13 mouse tissues

Kinase activity profiling was performed on 13 different mouse tissues to determine whether the similarities and differences in kinase profiles between the different tissues types. 21-day-old male

Swiss Webster mice were sacrificed by CO_2 asphyxiation after overnight feeding. Tissues from three male mice were harvested, rinsed with saline solution and flash frozen using liquid nitrogen until further use. Frozen tissues from the three mice were combined, pulverised on dry ice, and homogenised in a Dounce homogeniser in the KAYAK buffer as described in Section 2.7.2.

The kinase activity profiles in 13 different mouse tissues (brain, lung, testis, kidney, brown fat, liver, pancreas, white adipose tissue, spleen, heart, tibialis, small intestine and large intestine) were determined using 20 µg of lysate and the 96-well-plate KAYAK format. For nine of the mouse tissues (brain, lung, testis, kidney, brown fat, liver, pancreas, spleen and heart) the kinase activity profiles were compared with the kinase abundance profiles (Figure 4.1) that had previously been determined using shotgun proteomics[310].

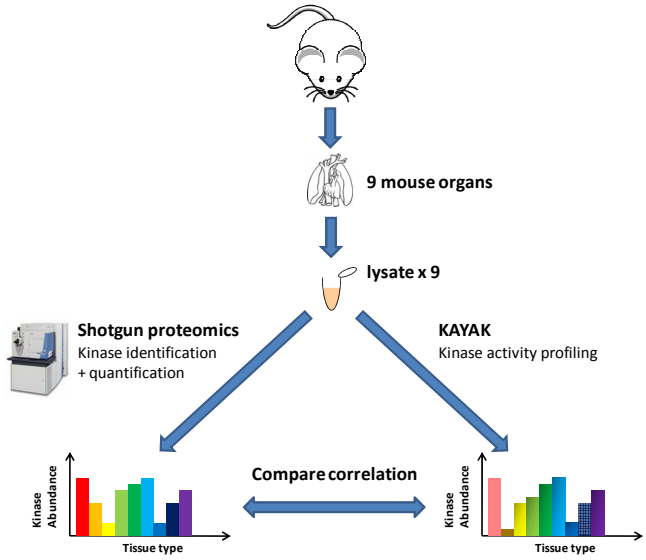


Figure 4.1 Workflow for kinase activity - kinase abundance profiling using 9 mouse tissues. Individual tissues from 3 mice were pooled and analysed for their kinase activities using KAYAK and their protein/kinase abundance using shotgun proteomics. Pearson correlation was used to compare the kinase activity/kinase abundance profiles.

4.2.2 Correlation profiling using kinase activity/kinase abundance

While the previous experiments measured kinase activities, Kinase activity/kinase abundance correlation profiling was also performed to determine which kinases were responsible for the phosphorylation of particular peptides. Protein separation using anion exchange chromatography was performed as described in Section 2.7.6. Kinase activity profiling was performed using the K60 substrate set and selected peptides from a large scale cell-cycle quantitative phosphorylation study[176]. KAYAK was performed with the selected ten peptides (K10) using HeLa nocodazole arrested and

asynchronous cell lysate (2.6.1). Only six of the peptides showed significant activity and were differentially active between the two cellular states (shown in blue in Table 4.1). To try to identify the kinases that are responsible for their hyper-phosphorylation during mitosis, these six peptides and two controls were utilised in the KAYAK assay. The two positive controls (B11 and C2) were chosen because of their high rate of phosphorylation in the nocodazole arrested state. The group of six peptides shown in blue together with B11 and C2 are termed the K8 substrate set.

Table 4.1 Peptides selected from mitotic phosphorylation study for analysis by KAYAK. The peptides highlighted in blue showed significant differences in phosphorylation in asynchronous and G2/M arrested states and were used for kinase activity/kinase abundance correlation profiling.

Peptide	Sequence	Description of the protein from which the peptide sequence is derived
5B10	EFAAALtAERIKTfK	EZH2, T300, enhancer of zeste homolog 2 (*Drosophila*)
5B11	HFRNVHsEDFENRfK	ADNP, activity-dependent neuroprotector
5B2	SDGAPAsDSKPGfK	BASP1, brain abundant, membrane attached signal protein 1
5B3	VEDAADsATKPENLfK	MKI67, antigen identified by monoclonal antibody Ki-67
5B4	KPSQVSsGQKLGPfK	BMS1, ribosome assembly protein (yeast)
5B5	EVDYSDsLTEKQWfK	SMARCA1, SWI/SNF related, matrix associated, actin dependent regulator
5B6	DDFKLNsSIVEPKfK	NFKB2, nuclear factor of kappa light polypeptide gene enhancer in B-cells
5B7	IEKLNSsLHFLQQfK	CAMSAP1L1, calmodulin regulated spectrin-associated protein 1-like 1
5B8	AQHPDYsFGELSRfK	PB1, polybromo 1
5B9	KTLENQsHETLERfK	NES, nestin

4.3 Results and discussion

4.3.1 Optimisation of a mass spectrometry based kinase assay for kinome profiling

Phosphorylation of a substrate by a kinase is highly dependent on the local amino acid sequence flanking the phospho-acceptor site, known as the 'phospho motif'. For example, AKT is known to phosphorylate peptides containing the motif: RXRXX[pS/pT][311]. Various factors are also important, including the activity state of the kinase, substrate availability and configuration[312]. With the use of peptide library approaches, the preferred motifs for many kinases have been identified in literature studies.

Designing a substrate peptide that is specific for a single kinase is a challenging task. The K90 set was developed by the Gygi laboratory to include substrates for kinases in various signalling pathways, as well as peptides that had been identified from large scale phosphoproteomic studies for which the kinase was not known. In the K90 set, there is a large amount of redundancy of the peptides. To reduce the redundancy and increase the number of pathways covered, many peptides with diverse sequences were synthesised, purified, and analysed with various cell lysates to identify peptides which show high phosphorylation rates with high specificity. From a library of approximately 300 peptides, 60 were selected, 28 of which were from the previous K90 selection. To assess these 60 peptides, the basal kinase activity in 12 different human cell lines was profiled with K60 using a substrate concentration of 5 μM and 1 μM. Kinase activity in a human cell line (HEK 293 T) in which certain kinases were overexpressed was also performed.

4.3.1.1 Kinase profiling of 12 human cell lines

Kinase profiling of the basal activities in the different cell lines can elucidate the activation state of signalling pathways and even identify hyperactivated kinases. KAYAK using the K60 substrate set was performed to detect signature kinase profiles for the 12 human cell lines, (Figure 4.2). Hierarchical cluster analysis was performed on both the peptide substrates and the cell lines to group similar kinase activity profiles and similar cell lines together.

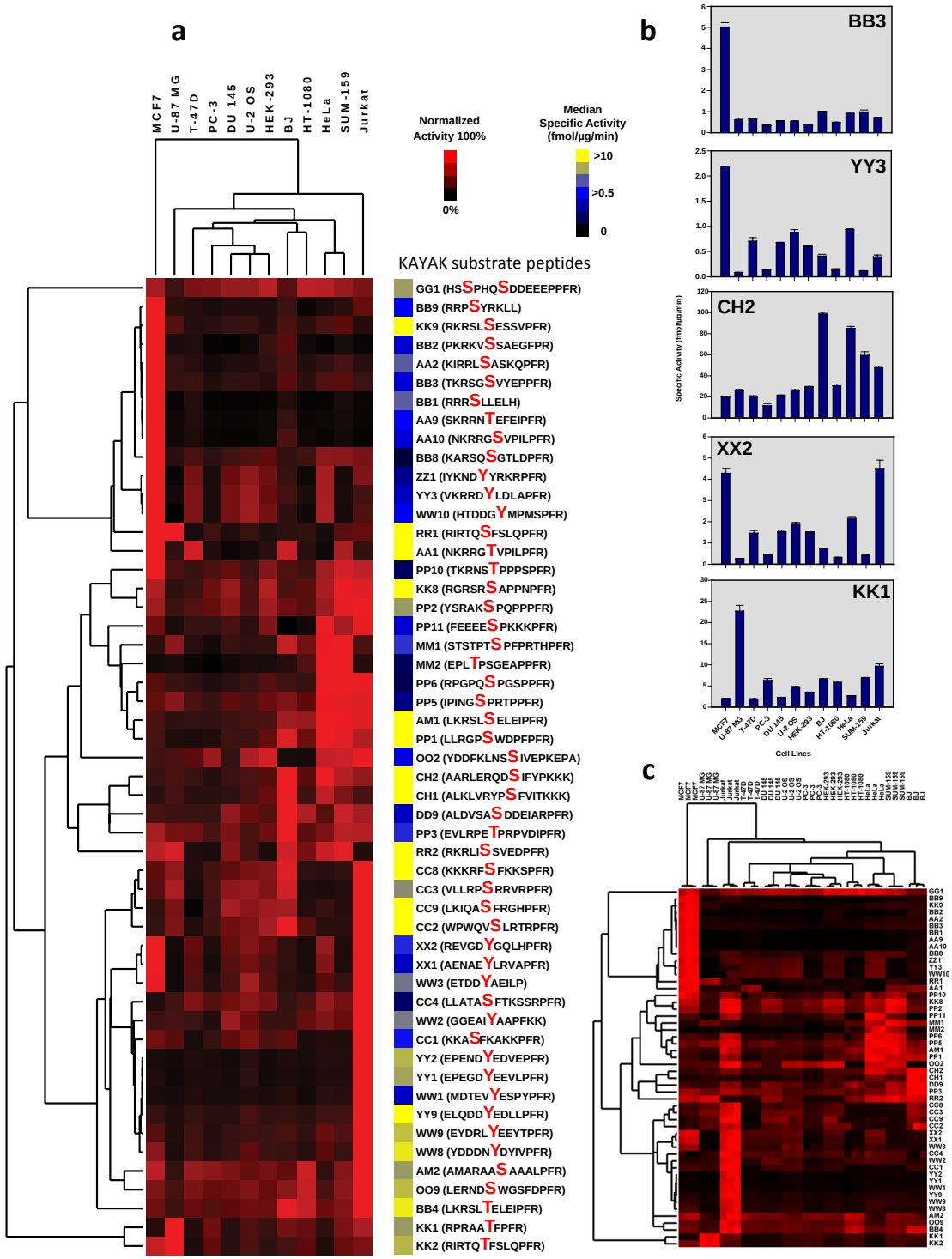


Figure 4.2 Profiles of kinase activities across 12 human cell lines using the K60 peptide substrates. The 12 cell lines comprise: MCF7 (breast), U-87 MG (glioblastoma), PC-3 (prostate), DU 145 (prostate), U-2 OS (osteosarcoma), HEK-293, T-47D (breast), BJ (foreskin fibroblast), HeLa (cervical), SUM-159 (breast), HT-1080 (fibrosarcoma), Jurkat (T lymphocyte). (a) Heat map of kinase activities using 5 µM peptide substrate concentration where the activity was normalised to the median value in each row, followed by hierarchical cluster analysis. b) Examples of individual peptide profiles from (a). Kinase activity values are the mean value from triplicate analyses with the error bars indicating the standard deviation. c) Profiles of kinase activities showing individual replicates. The three technical replicates for each of the cell lines clearly cluster together.

The two prostate cancer cell lines (PC-3 and DU 145) cluster together and show decreased MAPK substrate peptide (MM*) activities. Whilst the three breast cancer cell lines (MCF7, T-47D and SUM-159) do not cluster together and appear to have different profiles, the PKC activity (CC peptides) appears downregulated in all three samples. Interestingly, peptides designed for similar kinase targets are also observed to cluster together. For example, the PKC peptides are largely grouped and KK1 and KK2, which are known targets of AKT, are also clustered together. There is a further cluster of proline directed (PP) and MAPK (MM) substrates which have high phosphorylation in HeLa, SUM-159 and Jurkat lysates. Jurkat cells show significantly higher phosphorylation than the other cell types for at least half of the peptides, particularly the tyrosine kinases (WW, XX, YY, ZZ peptides) and the PKCs. MCF7 cells showed higher phosphorylation for PKA substrate peptides (AA peptides) and many basic peptides (BB). U-87 MG displays extremely high AKT activity (KK1 and KK2). This is consistent with the fact that the U-87MG cell line has a PTEN mutation, resulting in upregulation of PI3K and subsequent AKT activation. The same pathways that were shown to be upregulated in the K90 substrate set are also observed in the K60 substrate set, illustrating that the K60 set behaves similarly to the K90 and no major loss of pathway coverage is observed.

The K60 substrate set also covers more pathways than the K90 substrate set as it includes peptides that are substrates of checkpoint kinases. Two such substrate peptides which are included in K60 but not K90 are the peptide CH1, which is a substrate of CHEK1, and CH2, a putative substrate of CHEK2. Both of these peptides demonstrate high activity in BJ sells. BJ foreskin fibroblasts display high antioxidant capacity and slow telomere shortening[313]. BJ cells and HEK293 cells are the only non-cancer cell lines included in the study. It is possible that the BJ cells minimise the accumulation of genomic damage by a hyperactivated DNA damage response. It is known that stress induced senescence and telomere-dependent replicative senescence processes result in the activation of the DNA damage response kinases (ATM, ATR) and the downstream effector kinases CHEK1 and CHEK2[314-316]. The kinase activities of the checkpoint kinases are fairly low in the cancer cell lines, indicating that the DNA damage repair pathway might be suppressed. It would be interesting to compare the kinase activity of the checkpoint kinases in other human fibroblast cell lines to determine whether high CHEK1 and CHEK2 activity is also observed. Other kinases included in the K60 set and not in the K90 include GSK3 and FAK. The peptides GG1 and GG2, proposed GSK3 substrates, appear to be fairly uniformly phosphorylated across the 12 cell lines except with lower phosphorylation observed in U-87 MG and BJ cells. WW3 is thought to be a substrate of FAK and is highly phosphorylated in Jurkat and MCF7 cells. These examples demonstrate the broader coverage the K60 set provides and its potential implications for understanding of intracellular protein activity dynamics.

KAYAK assays were performed with three replicates. Very little variation is seen between the different replicates as can be observed by the small standard deviations (shown as the error bars in Figure 4.2b) and the clustering of each of the three replicates (Figure 4.2c). This suggests that the KAYAK technique is highly reproducible and generates very robust results.

While profiling was performed at 5 μM for the K90 substrate set[150], lower substrate peptide concentrations may have been feasible. Therefore, upon conducting profiling experiments utilising the K60 set, concentrations of both 5 μM and 1 μM were tested. The kinase activity profiles at the two concentrations were very similar (Figure 4.3). However, when comparing the columns for each cell line at 5 μM and 1 μM, a few differences can be observed. For example, for BJ at 5 μM peptides BB2, BB1, AA10 and AA9 all show significant phosphorylation, but at the 1 μM concentration the phosphorylation of AA10 becomes much more significant. Also, at 1 μM the lower GSK3 activity (GG1 and GG2) is much more significant than at the 5 μM concentration. In the PC-3 cell line, WW3 has significant phosphorylation which was not observed at the 5 μM concentration. The cluster of PKA/basophilic kinase activity (AA2 to BB8) appears very similar, with the activity in the MCF7 cell line dominating the others. However, with the lower substrate concentration, more differences were observed between the different peptide profiles. It might have been expected that at the lower substrate concentration (1 μM) the activity of many of the peptides could be below detection levels, but this was not the case, and suggests that a lower concentration could be used routinely. Lowering the substrate concentration would be expected to lead to increased substrate specificity and reduced kinase promiscuity, and there is some evidence of this for a few peptides where the profile changes between the two concentrations, but in general the profiles were very similar.

All the kinase activities detected by the K60 substrate set were still detectable at 1 μM and the general similarity of the protein profiles at both concentrations indicates that the next generation KAYAK assay could be based on a lower substrate concentration of 1 μM. Advantages of using lower substrate concentrations include the reduced peptide consumption (lower cost) leading to significant time saving in peptide purification, reduced sample clean-up procedures prior to mass spectrometric analysis as well as the potential to expand the number of peptides included to perhaps 500.

It is important to consider that although protease inhibitors are present in the cell lysate, degradation or modification of the kinase substrate peptides or the phosphorylated form may occur by proteases or other enzymes that are not inhibited. This may result in an underestimation of the degree of phosphorylation observed. The extent of possible degradation and modification of the substrate peptide will be investigated in future work.

In the future, kinase activity profiles may be useful in the investigation of drug effects (e.g. kinase inhibitors) in different cell types. There is much less redundancy in the K60 substrate set than in the previous K90. Whilst some redundancy can still be observed (for example with peptides CC8 and CC3, XX1 and XX2), the peptides may react differently under different circumstances and a core number of peptides against certain kinases is needed. For example different CC peptides have preference for the different PKC isoforms. To test a substrate set thoroughly it is necessary to profile a large number of different samples to determine the usefulness of each of the individual peptides.

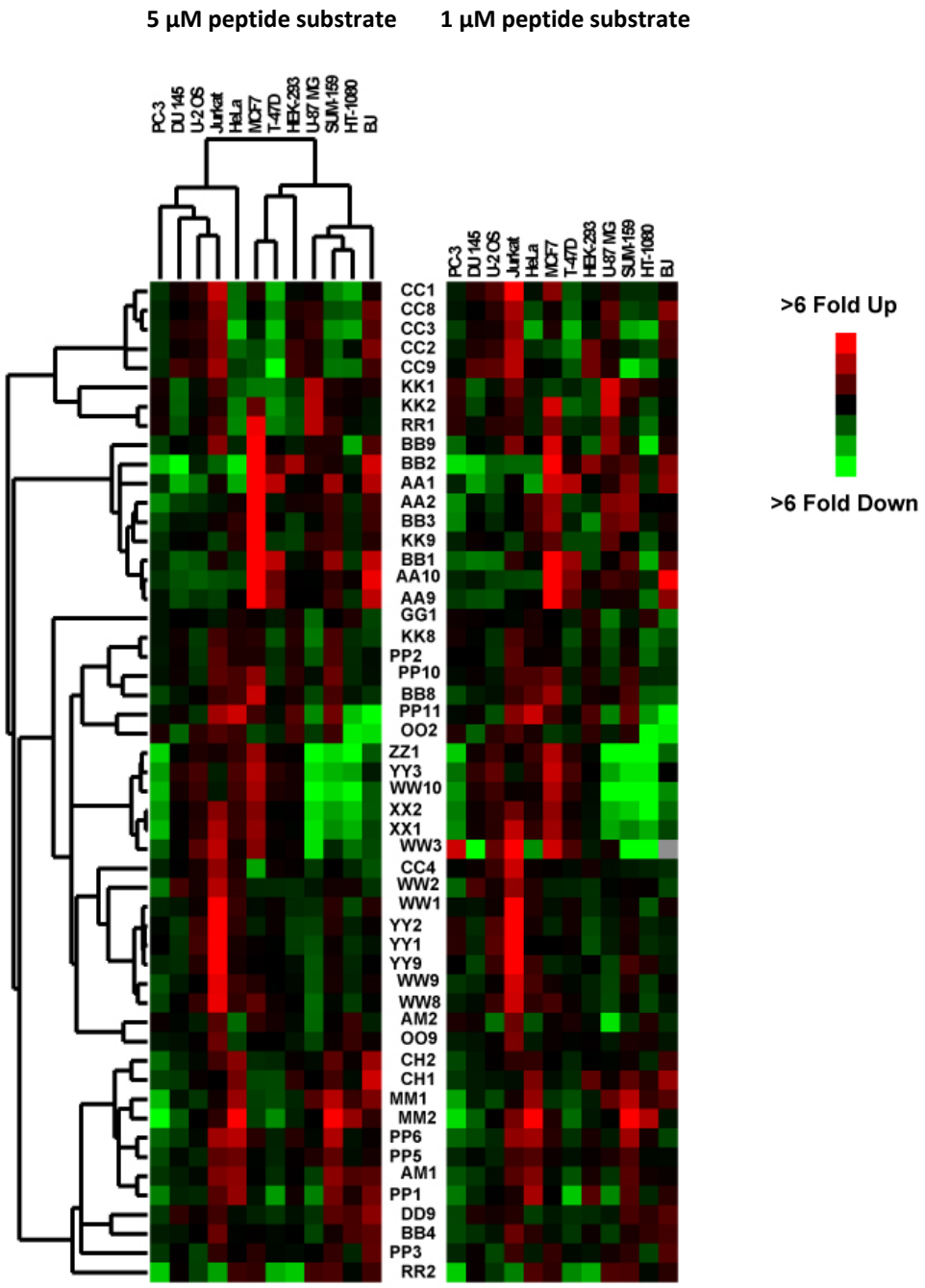


Figure 4.3 Profiles of kinase activities across 12 different human cell lines at 5 µM and 1 µM. Measurements were performed in triplicate and the mean value of the kinase activity is shown for each peptide for each cell type. The activity is normalised to the median value of the row, followed by hierarchical cluster analysis. The 12 cell lines comprise: MCF7 (breast), U-87 MG (glioblastoma), PC-3 (prostate), DU 145 (prostate), U-2 OS (osteosarcoma), HEK-293, T-47D (breast), BJ (foreskin fibroblast), HeLa (cervical), SUM-159 (breast), HT-1080 (fibrosarcoma), Jurkat (T lymphocyte).

4.3.1.2 Kinase profiling of a human cell line with overexpressed kinases

One possible method for associating peptide substrates and the kinases responsible for their phosphorylation is the use of recombinant kinases. Disadvantages of such an approach include kinase promiscuity and the fact that kinases may not be in an active configuration (may need cofactors, binding partners). Using a kinase overexpressed in a cell lysate avoids many of these problems and may give a more accurate representation of phosphorylation activities *in vivo* as protein-protein interactions are maintained. Purified kinases lack cellular context and results obtained in this way may be misleading**.**

These initial studies involving the overexpression of 11 kinases (Figure 4.4a and b) are part of a larger project goal to overexpress and knock down a large library of kinases. This project aims to help elucidate the kinase substrate specificities, as well as investigate the effect on kinase activity by perturbing kinase signalling pathways.

Notably, overexpression of particular kinases in the HEK 293 T-cell line lead to greater phosphorylation of expected substrates in the majority of cases. AKT1 overexpression lead to increased phosphorylation of KK1 and KK2 (known AKT substrates) by over 5 fold (Figure 4.4c and d). Overexpression of PRKCB1 (PKC beta) leads to overexpression of a cluster of PKC target peptides (CC8, CC1, CC3, CC9). Interestingly, overexpression of PRKCZ (PKC zeta) only displays significant phosphorylation of CC9 and not the other PKC peptides, indicating that the peptides are specific to different PKC isoforms. MAPK14 leads to increased phosphorylation of many proline directed and MAPK target peptides (MM1, PP3, PP7, PP6, MM2, PP9) as would be expected. PAK2, TOPK and the control have very similar profiles, indicating either that these kinases were not active or there are no peptides in the K60 which are able to report on their activity. This proof-of-concept work sets a foundation for a larger study involving the overexpression of a large number of kinases.

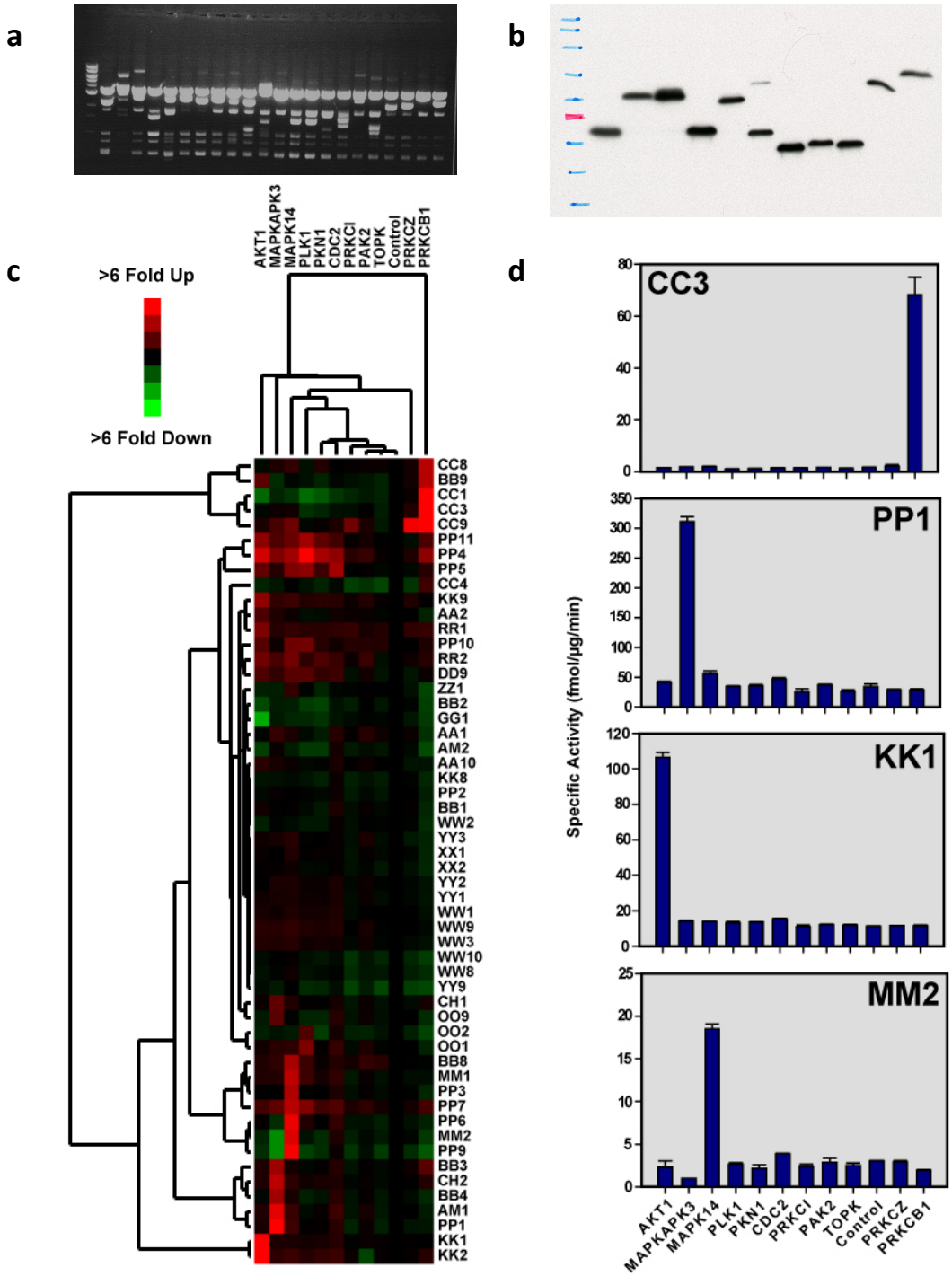


Figure 4.4 Overexpression of kinases in HEK 293T cells using Gateway cloning. (a) DNA digests on 1% agarose gel. From LHS to RHS: Std, AKT1, CDK7, CHEK2, MAPK14, MAPK6, MAPKAPK3, PAK2, PKN1, PLK1, PLK2, PRKAA1, PRKCB1, PRKCD, PRKCG, PRKCH, PRKCI, PRKCZ, PBK, TP53RK, CDC2, PRKAA2, AURKB. (b) Western blot analysis of 11 lysates using HA antibody. From left to right: AKT1, MAPK14, MAPKAPK3, PAK2, PKN1, PLK1, PRKCB1, PRKCI, PRKCZ, TOPK, CDC2. Only 11 of the 22 kinases overexpressed were expressed according to western blotting. (c) Heatmap of kinase activities for HEK 293 T cells with 11 overexpressed kinases. Activities were normalised to the control cell line in which no kinase was overexpressed. Hierarchical cluster analysis was performed which groups peptides and cell lines with similar activities together. (d) Examples of several peptides from c.

4.3.2 Applications of mass spectrometry based kinase assay for kinome profiling

This section presents the results obtained from kinase activity profiling performed on a range of samples. Activity profiles were obtained for several organisms (human cell culture, mouse, *Drosophila*, yeast), human saliva and 13 mouse tissues.

4.3.2.1 Kinase activity profiles in different organisms and tissue types

To determine whether kinase profiling using KAYAK would be useful in organisms apart from humans, the method was assessed on lysates from yeast (Appendix K) as well as homogenates from different mouse tissues (Appendix J) and *Drosophila* embryos (Appendix I). Although the peptide substrates had been designed for human kinases, significant kinase activity was detected in the various organisms against many of the peptides in the K60 substrate set (Figure 4.5). Furthermore, for various peptides, activity was greater in organism extracts than some human cell lysates. The results from the samples cluster together based on the organism from which they are derived. There are two very distinct branches: one for mouse and human and the other for *Drosophila* and yeast. This is expected because human and mouse are more closely related to each other than to *Drosophila* and yeast.

PKC substrate peptides were highly phosphorylated by mouse liver extract and many proline directed substrate peptides were significantly phosphorylated by yeast extract. Proline directed kinases are involved in cell cycle control, and for this reason it might be expected that they are conserved from yeast to humans since cell cycle control is such a fundamental process. *Drosophila* embryo extracts displayed significant phosphorylation of a range of substrates, but very few tyrosine substrate peptides were phosphorylated. Whilst some tyrosine phosphorylation was observed in mouse, the majority of the tyrosine kinase substrates were most highly phosphorylated in human HEK cells. This type of comparison could indicate which kinase motifs are conserved through evolution. Phosphorylation of peptide substrates for PKAs, PKCs and MAPKs are observed for the four organisms. However, significant phosphorylation of the tyrosine kinase substrates was only observed in humans and to a lesser extent in mouse.

Due to the many gains and losses of kinase groups during evolution, the distribution of kinase families is very different across the major phyla. The emergence of the metazoa from opisthokonts resulted in 76 new classes of kinases including, most importantly, the tyrosine kinase group[317]. Yeast, a fungus and earlier branch on the opisthokont line, does not have any conventional tyrosine kinases, explaining the complete absence of the phosphorylation of any tyrosine kinase substrates in yeast (Appendix K).

The presence of kinase activities in the different organisms confirms that KAYAK can be used successfully to profile kinase activities in organisms other than human.

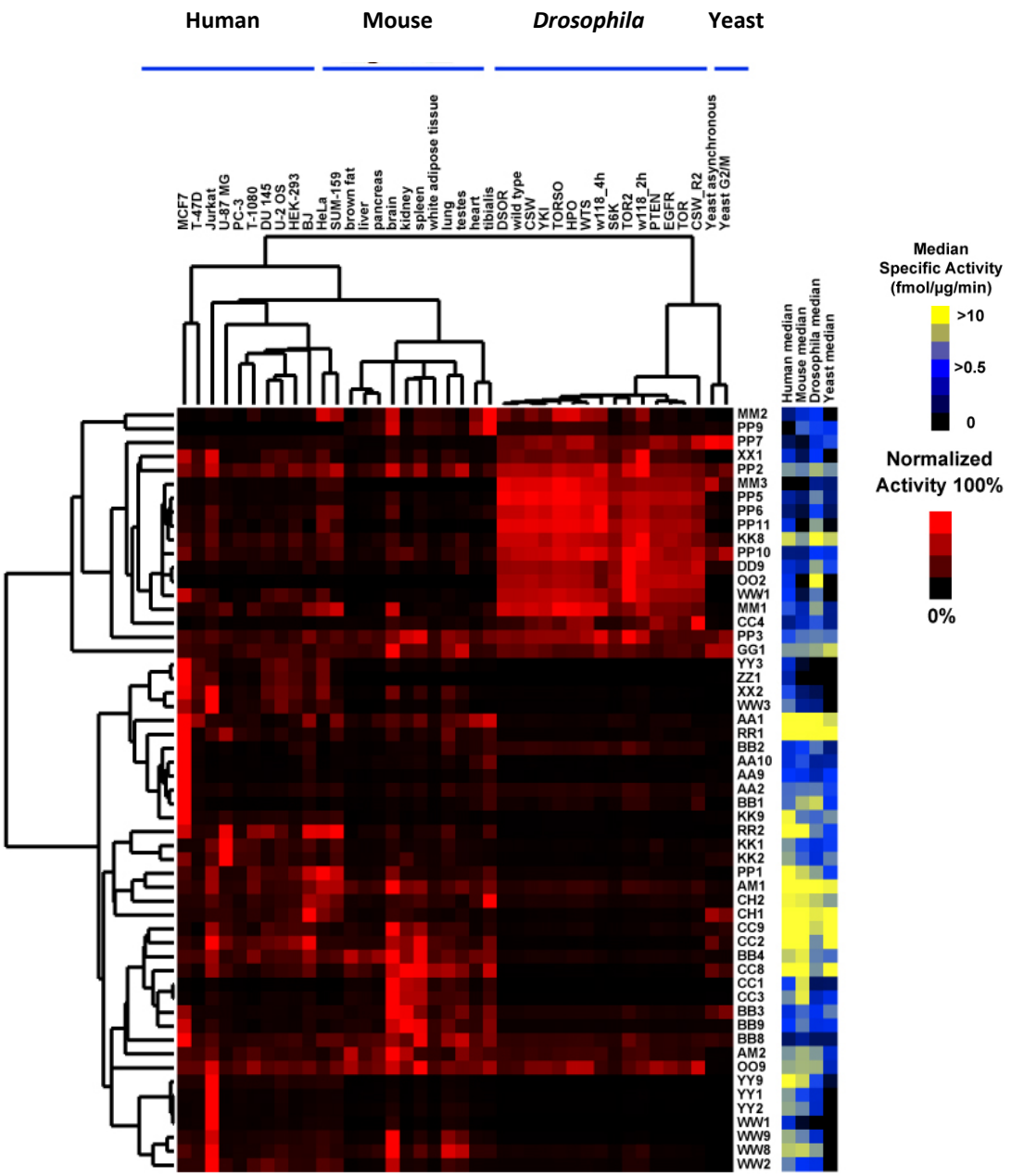


Figure 4.5 Profiles of kinase activities across four different organisms: human, mouse, *Drosophila* and yeast. KAYAK was performed in triplicate using 20 µg of lysate from each sample. The activity is normalised to the maximum value of the row. Hierarchical clustering was used to group similar responding peptides together. In total, 12 human cell lines, 11 different tissues from mouse, 15 samples from *Drosophila* embryos and 2 samples from yeast were analysed using KAYAK.

4.3.2.2 Kinase activity profiles in saliva

Kinase activity profiling using the K60 substrate set was performed on the saliva supernatant and saliva residue from 6 subjects (Figure 4.6a), and significant kinase activity was observed in both the saliva supernatant and residue.

Saliva supernatant shows significant phosphorylation of peptides AA1, AA9, AA10, very similar to the activities observed for the saliva residue, indicating high PKA activity. Significant phosphorylation of RR1 and RR2, which are known RSK substrates, as well as for GG, a substrate for GSK3, and AM1 a substrate from AMPK was observed. High activity in the residue was also observed for PKC substrate peptides CC1, CC8, CC3, CC9, and CC2, with slightly lower levels observed in the supernatant. Significant basophilic kinase activity was also present in the residue with BB2, BB1, BB9, BB3, and BB4 being highly phosphorylated. MAPK activity was detected in the residue but very little was observed in the supernatant. In general, very low tyrosine activities were observed in both the residue and the supernatant. Whilst greater kinase activity was observed on average in the residue compared to the supernatant, there were several peptides that had similar or even sometimes greater activities in the residue.

Large inter-subject variation in kinase activities was clearly seen in Figure 4.6b. One may expect that if, for each subject, supernatant and residue kinase activity profiles are similar, this result may be an artefact of dead epithelial cells. However, the kinase profiles of the lysed cells may be different due to protein/kinase degradation or due to the different signalling pathways involved in apoptosis and necrosis. In general, for the same amount of total protein measured in each (10 μg), greater KAYAK activity was observed in the saliva residue when compared to the supernatant.

To try to determine the source of the kinase activity in saliva, the kinase activities obtained in the saliva supernatant were compared with those in the residue. The bulk of the kinase activity in the supernatant was expected to originate from lysed cells. To determine whether glandular secretions contribute to kinase activity KAYAK will be performed on gingival crevicular fluid and pure glandular secretions in the future.

The role of the observed kinase activity in saliva has not been studied previously, and could be of potential biological interest. ATP (source of phosphate for kinase reactions) is known to be present in saliva at concentrations ranging from 8 to 1515 nM, and has been shown to be directly correlated with the bacterial count and epithelial cell numbers[318]. Of the total ATP concentration in whole saliva, it was shown that approximately 14% of the ATP was in the soluble fraction and not in the human cellular or bacterial residual portion[318]. The source of this ATP in the soluble saliva fraction is not known, but is likely to be the release of ATP from dead mouth epithelial cells and bacteria. It is not known whether this ATP has a role in saliva.

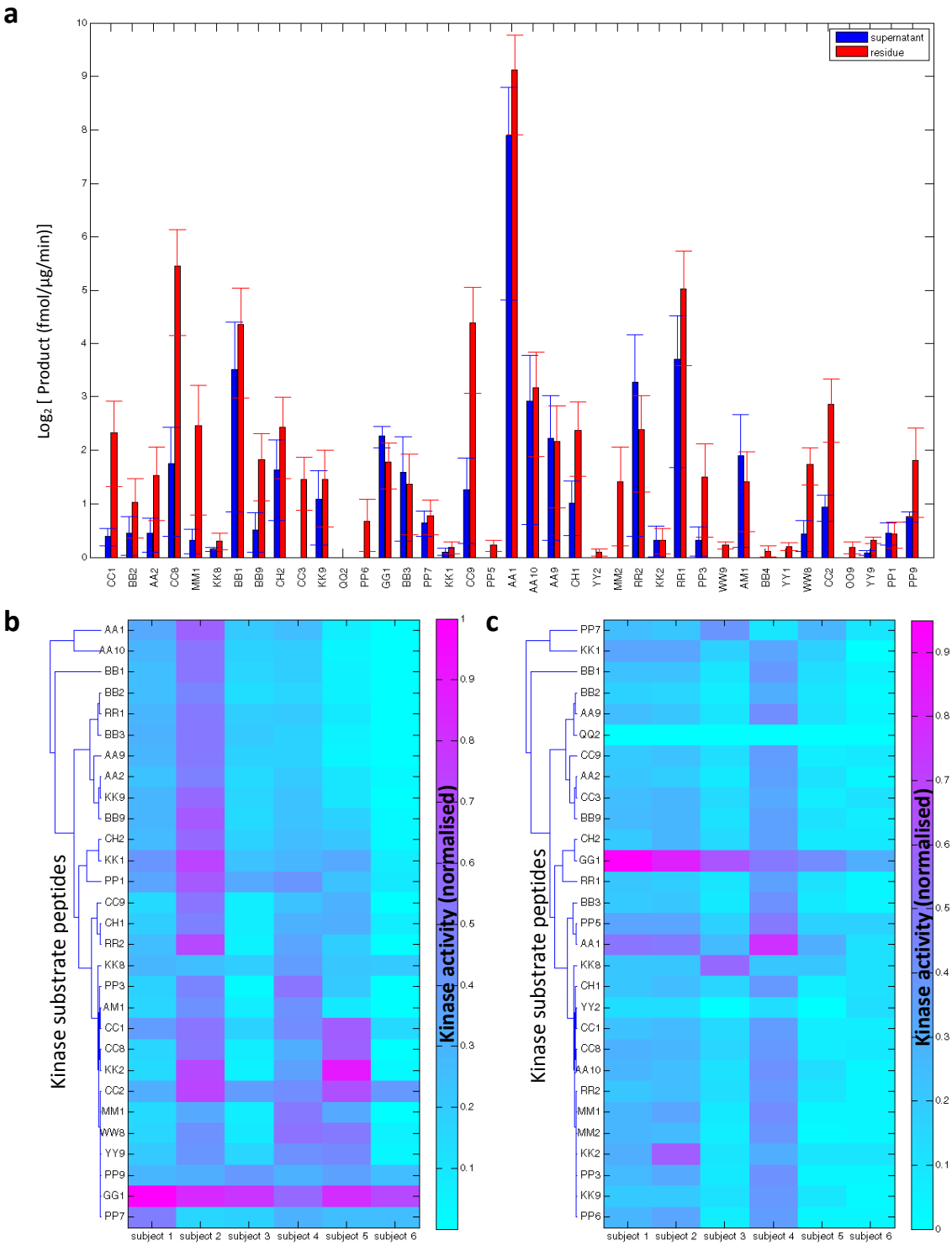


Figure 4.6 Profile of kinase activities in saliva supernatant and residue. (a) Average kinase activities for 6 subjects are compared between the saliva supernatant and residue. Error bars indicate the standard deviation between the 6 subjects. Kinase activities using saliva supernatant (10 µg) and residue (10 µg) were measured using the K60 substrate set. Heat map of KAYAK activities in saliva (c) supernatant and (b) residue for six people. Analysis by Pearson correlation hierarchical clustering was used to group similar responders together. Each row represents the phosphorylation rate of a each peptide normalised to the sum of kinase activities for the 6 subjects.

4.3.2.3 Kinase activity profiling in 13 mouse tissues

Whilst generalised cell signalling models and phosphorylation networks have been determined, these have been generated using model systems, but the signalling pathways and networks are likely to be different in various tissues and diseased states. Global characterisation of kinase activities in different tissues can ultimately aid in generating a picture of which pathways may be significant in particular tissue types. Kinase activity profiling was performed on 13 different tissues from a mouse: brain, lung, testis, kidney, brown fat, liver, pancreas, white adipose tissue, spleen, heart, tibialis (muscle), small intestine and large intestine (Figure 4.7).

The kinase profiles of the small and large intestines cluster together and are very different from the profiles of the other mouse tissues. For the small and large intestine extracts, several basophilic kinase substrates are highly phosphorylated (BB1, BB2, BB3 and BB9) as well as several PKC substrate peptides (CC1, CC2 and CC4) and CHEK kinase substrates (CH1, CH2). However, the activity of the tyrosine kinases are much lower than those in the other tissues. It is possible that natural bacterial flora present in the small and large intestine could be contributing to the kinase activity profile and could explain the very different kinase activity profiles observed in the intestinal tissues. Much less is known about bacterial kinases compared to those in eukaryotes, and whilst most prokaryotic signalling and protein phosphorylation is proposed to occur through histidine-aspartate kinases, there are various eukaryotic Protein Kinase-Like kinases (PKLs) known to be present in bacteria[319]. Whilst there are known to be tyrosine specific kinases present in bacteria, it might be expected that their motifs are less well conserved than for serine/threonine specific kinases since in eukaryotic evolution the tyrosine kinase class was only added with the emergence of metazoans[317]. Therefore, very little tyrosine phosphorylation by bacteria of the human tyrosine kinase substrates would be expected.

Another reason why the kinase activity profiles for the small and large intestine are so different from the other tissues could be on account of degradation or modification of the kinase substrate peptides (or phosphorylated product peptides) through the large number of proteases present in the small and large intestines. Although protease inhibitors are added to reduce proteases, the wide variety of proteases inevitably leads to some still remaining active. The small and large intestines are have a much larger proportion of proteases compared to the other tissues and this could explain and the lower kinase activities observed.

The results for the heart and the tibialis extracts cluster together, as would be expected for these two muscle tissues. PKA activity is high in both these tissues. This is consistent with mRNA profiling data where PKA abundance (PRKACA) is high in heart and muscle[320]. PKA is involved in the regulation of glycogen metabolism and occurs predominately in muscle tissue. With regard to the other profiled tissues, the greatest kinase activity is observed in the brain, including the phosphorylation of many tyrosine kinase substrates, followed by the lung, testis and spleen.

A comprehensive protein quantitation and phosphorylation study was performed using nine different

mouse tissue samples by post doctoral fellows in the Gygi laboratory[310]. Kinase activity profiling was performed using the same tissue samples that were used for protein quantitation (and four extra tissue types). This was done to complement the phosphomouse catalogue and determine whether changes in the kinase activity as assessed by KAYAK could be correlated with the abundance of kinases in the different tissues.

However, increased kinase abundance does not necessarily equate to increased kinase activity as the kinase may not be in its active state. The activation of many kinases involves the phosphorylation of specific activation loop residues which induces a conformational change. Taking this factor into consideration, post doctoral fellows in the Gygi laboratory mapped the spectral count of the activation loop phosphorylation sites on kinases as an indirect measure of kinase activity[310].

The kinase activity profiles of the nine mouse tissues, obtained using KAYAK (Figure 4.8a), were compared to the kinase spectral count abundance profiles (Figure 4.8b) and the spectral count of the kinase activation loop phosphorylation sites (Figure 4.8c). The kinase abundance profile is dominated predominately by the brain, with high kinase abundance also present in the lung, spleen and testis. For the kinase activity profiles, the brain also has the highest kinase activities. Since neural tissue, the primary constituent of the brain, contains a wide variety of different cell types and complicated signalling networks between the different cell types, high kinase activity would be expected and is consistent with the finding that the brain contains the greatest number of phosphopeptides[310].

The kinase activity profiles and the kinase abundance profiles were compared using Pearson correlation (Figure 4.9). Many of the PKC substrate peptides cluster together (CC1, CC3, CC4 and CC9) and have good correlation coefficients with PRKCD abundance (Figure 4.9b). RR1 and RR2 (RSK substrate peptides) have high correlation coefficients with RSK abundance, RPS6KA1 (Figure Figure 4.9c). Looking at specific kinase-substrate pairs, the correlation coefficient between AKT1 and KK1 is 0.71 with the lung displaying highest AKT1 abundance and greatest KK1 phosphorylation (Figure 4.10). Other examples include CDK5 and PP5 with a correlation coefficient of 0.93, with the abundance and activitiy most prominent in the brain. The abundance of casein kinase and phosphorylation of DD9, a substrate of casein kinase are highly correlated (Pearson correlation coefficient 0.98) with the maximum activity observed in the testes followed by the brain and then the spleen.

The kinase activity profiles were also compared with the spectral count of the phosphorylation sites on kinase activation loops (Appendix L). In general, the correlation was poorer than using the kinase abundance spectral count. This is likely to be on account of the use of different samples for the phosophoprotein identification and for the kinase activity assay. Whilst the same type of mouse was used for both, replicates from the same extracts were used for kinase activity and kinase abundance, and samples from different mice were used for the phosphorylation studies.

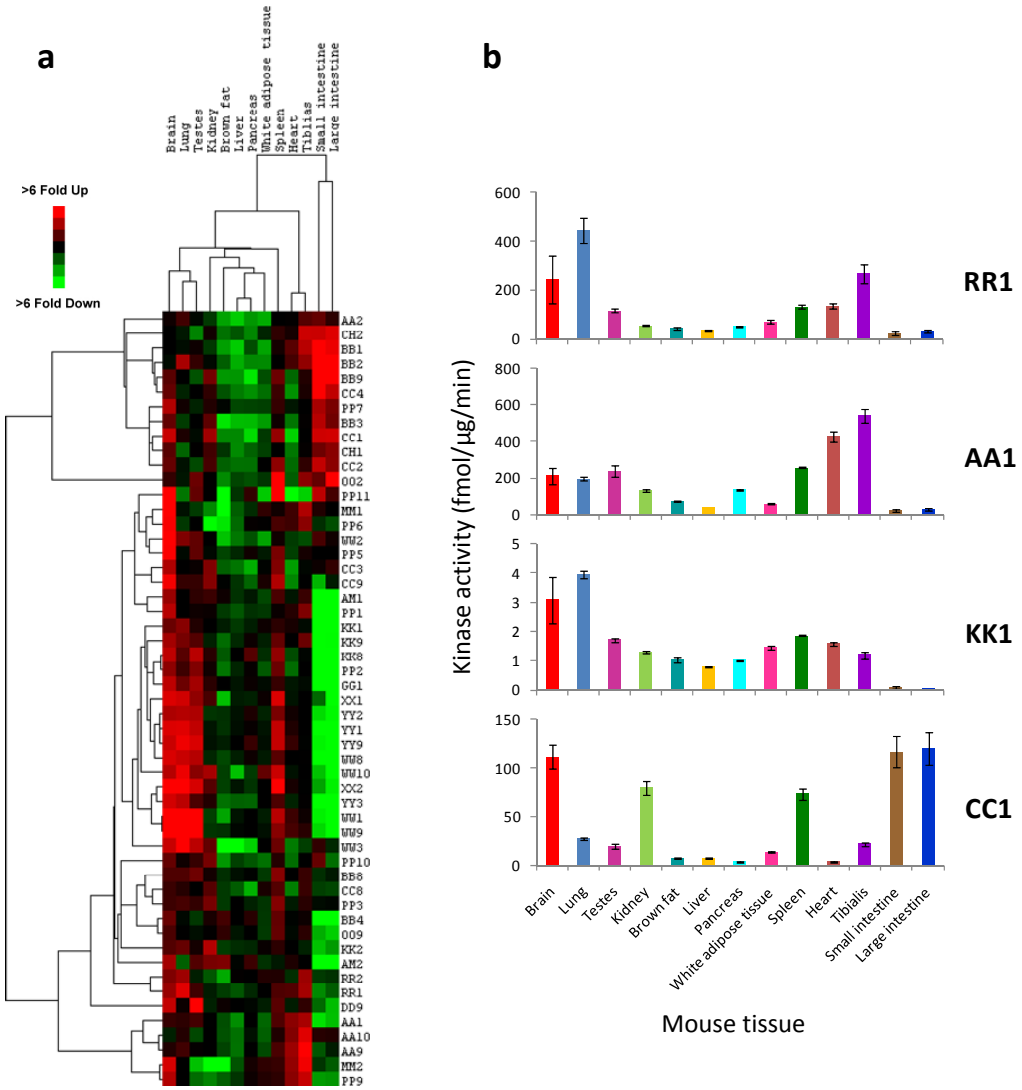


Figure 4.7 Profile of kinase activities in 13 different mouse tissues. (a) Heat map of kinase activities for mouse tissues from the brain, lung, testes, kidney, brown fat, liver, pancreas, white adipose tissue, spleen, heart, tibialis, small intestine and large intestine. Analysis by Pearson correlation hierarchical clustering was used to group similar responders together. Each row represents the phosphorylation rate of a particular peptide normalised to the median of kinase activities for each tissue. A maximum fold cut off of 3 was used. (b) Examples of kinase activities for particular peptides from (a). Error bars indicate the standard deviation of 3 replicated kinase reactions.

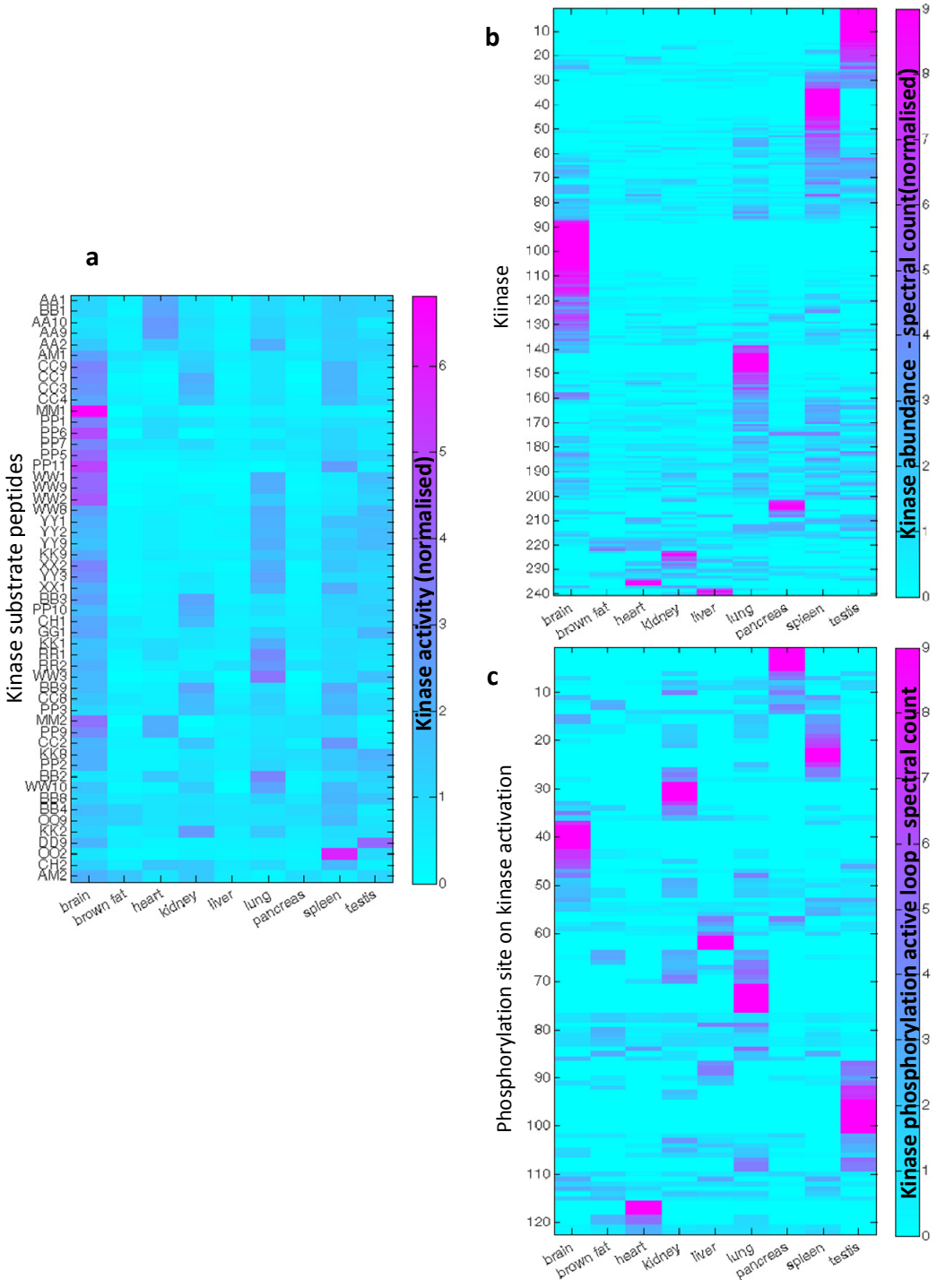


Figure 4.8 Profiling experiment with 9 mouse tissues. (a) Kinase activity profiles of 9 mouse tissues: brain, brown fat, heart, kidney, liver, lung, pancreas, spleen and testes using KAYAK. The kinase activities are normalised to the sum of all the kinase activities for a specific peptide across the 9 tissues and similar responding peptide substrates are clustered together by hierarchical clustering. (b) Kinase abundance profile in 9 tissues obtained using shotgun proteomics with spectral count as a measure of protein abundance. The abundance of the kinase is normalised to the sum of spectral counts for a specific protein across the 9 tissues and similar kinase abundance profiles are clustered together by hierarchical clustering. (c) Profiles of phosphorylation site spectral counts on the kinase activation loop across 9 mouse tissues. The abundance of the phosphorylation site is normalised to the sum of spectral counts for that site across the 9 tissues and similar profiles are clustered together by hierarchical clustering.

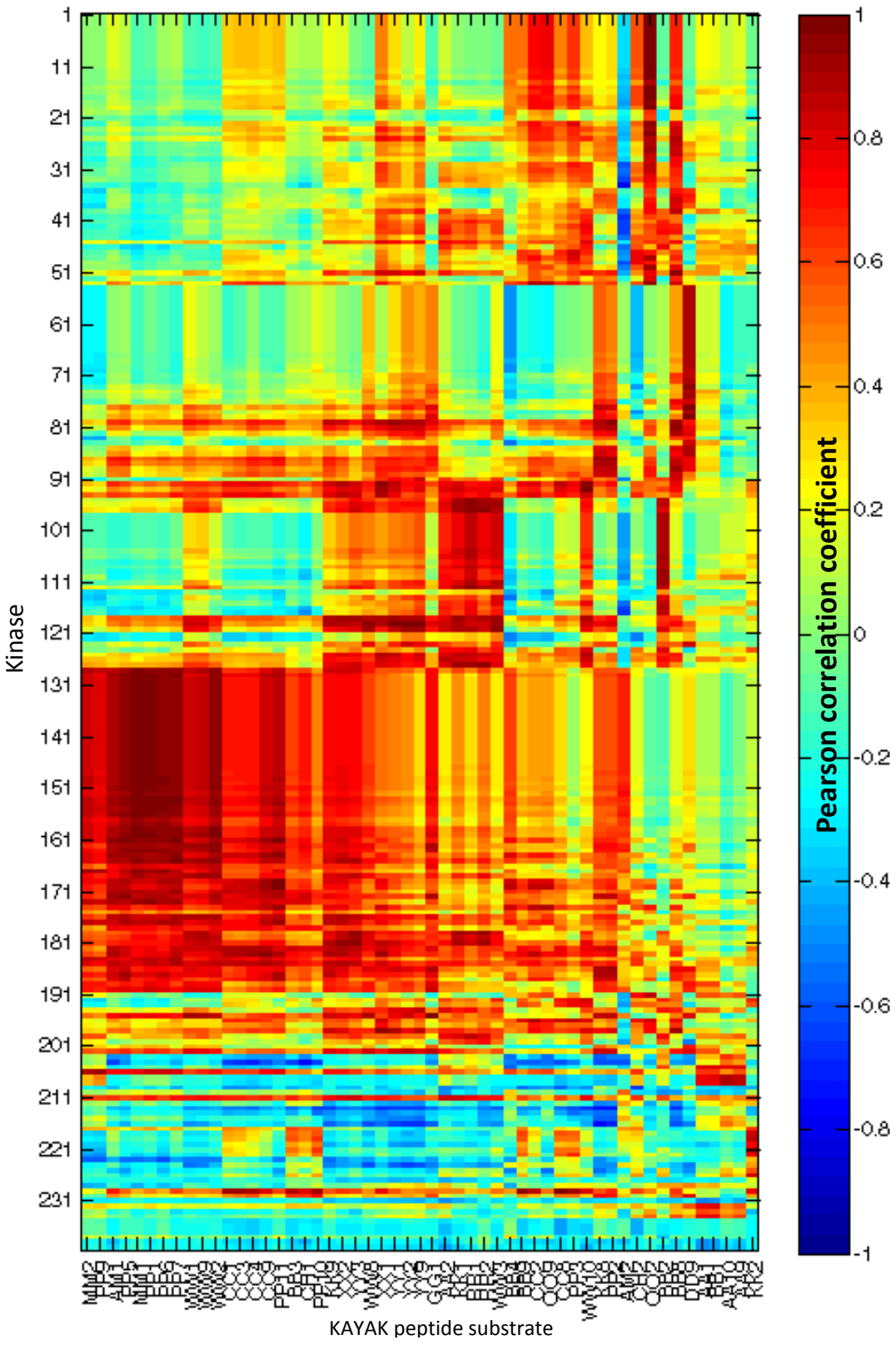


Figure 4.9 Correlation profiling for kinase activity-kinase spectral count. Heat map of Pearson correlation coefficients between the Kinase activity profile (KAYAK) and kinase abundance profile (spectral count) for the 9 mouse tissues. A Pearson correlation coefficient of 1 (shown in red) indicates a high degree of similarity between the profile of kinase activities in the different tissues and the kinase abundance.

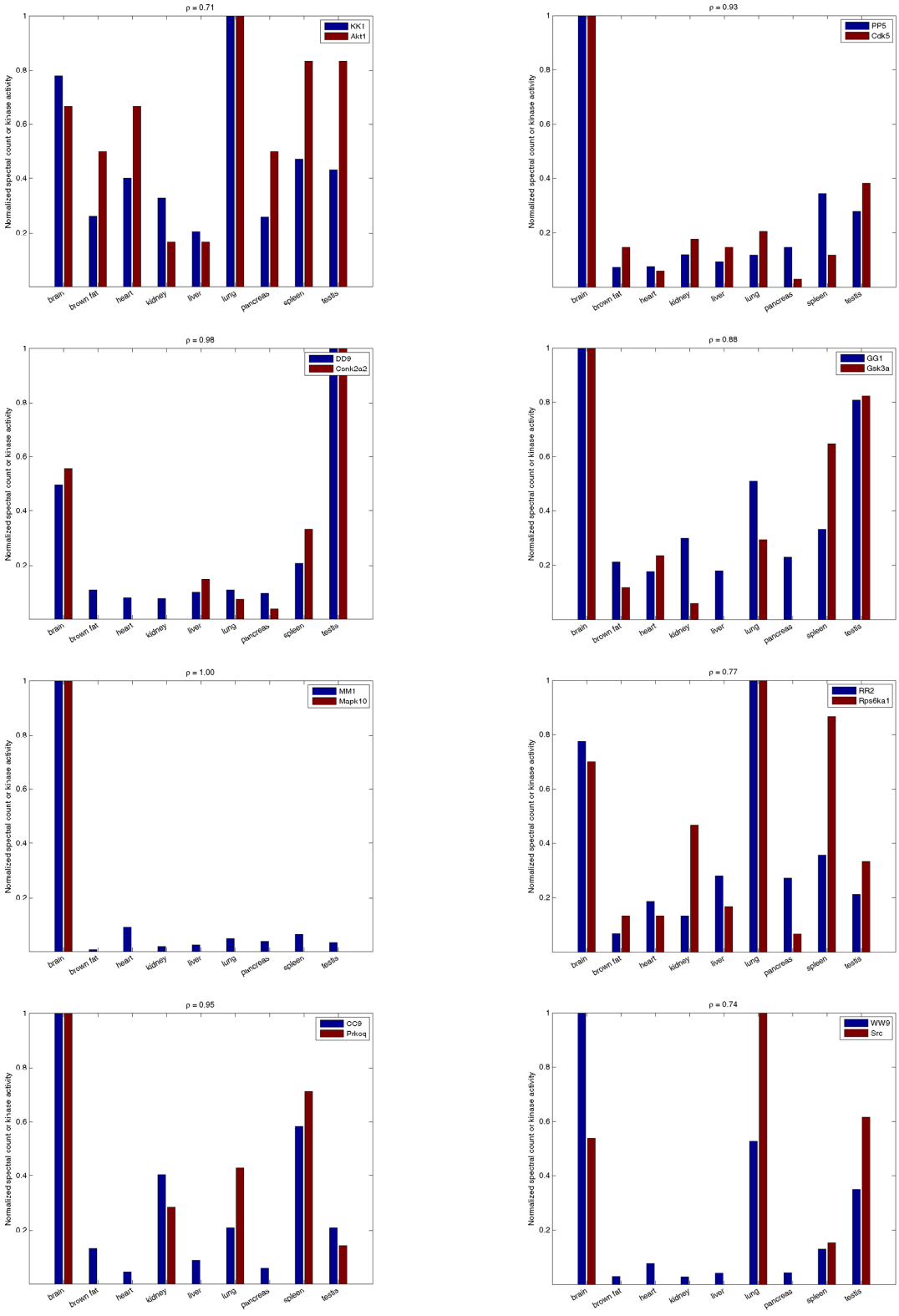


Figure 4.10 Correlation profile bar charts of several kinase-substrate pairs. The kinase acitivity and spectral count are each normalised to the maximum value observed across all tissues. The normalised spectral count and KAYAK activities are plotted for each tissue type. The Pearson correlation coefficient (ρ) for each kinase-substrate pair is given above each plot.

4.3.3 Correlation of kinase activity with kinase abundance to identify kinase-substrate pairs

Arguably the most important issue in phosphorylation studies is to determine kinase-substrate relationships. Kinase activity correlation profiling is a technique that can be used to identify the kinase responsible for a particular phosphorylation event and involves the fractionation of a lysate at the protein level. After the fractionation, each of the fractions is analysed for activity and an activity profile for a particular substrate peptide is obtained. This activity profile is compared with the kinase abundance profiles (obtained using shotgun proteomics) using Pearson correlation to obtain candidate kinases that could be responsible for the phosphorylation of the substrate peptide. Correlation profiling using kinase activity/kinase abundance, developed by Dr. Kubota in the Gygi laboratory, has previously been shown to be successful using anion exchange chromatographic separation, however, this technique may be applied to many other separation techniques. The use of gels as a means of chromatographic separation has many advantages on account of their ease of use and the feasibility of using a blue native gel separation in conjunction with an in-gel kinase activity/kinase abundance correlation profiling was tested. In-gel kinase activity/kinase abundance correlation profiling was previously described in Section 2.7.7. This was the first time that KAYAK had been performed in-gel and the proof of concept demonstrated significant kinase activity was observed in the gel. Whilst KAYAK in-gel correlation profiling was successful to a degree, and the principle works, the procedure suffers from the low resolution on the BN PAGE and is more suited to the analysis of simple samples. For more reliable kinase-substrate identification with complex samples, protein separation using anion exchange chromatography (AEX) is the preferred technique.

Kinase activity/kinase abundance correlation profiling using AEX was performed on HeLa cell nocodazole arrested cell lysate to identify kinase-substrate pairs. HeLa cell nocodazole arrested cell lysate was subdivided into 40 fractions using anion exchange chromatography. Each fraction was analysed by KAYAK using K60 substrate peptides to generate kinase activity profiles as well as by LC-MSMS shotgun proteomics to identify and quantitate the proteins and kinases present in the lysate. In total 5,626 proteins, including 172 kinases, were identified by shotgun proteomics. Hierarchical clustering was used to group peptide substrates with similar activity profiles together across the various fractions (Figure 4.11a). The majority of kinase activities for the different peptide substrates peaked between fractions 8 and 30, with low activity being observed in the early and late fractions. Similarly, the majority of the kinase abundance profiles had their maximum between fractions 8 and 30 (Figure 4.11b).

The kinase activity-protein abundance profiles were compared using Pearson correlation (Figure 4.12). Hierarchical clustering of the Pearson correlation coefficients indicates which peptide substrates may be phosphorylated by similar kinases or by kinases that are present in the same protein complex. Many of the substrate peptides for MAPK and proline directed kinases are clustered together (PP4, MM3, PP6, PP7, PP5, PP1, MM1, MM2, PP9, PP10, PP3). A cluster of PKC substrate peptides are

also present (CC1, CC3, CC8) as well as a large cluster of substrates for PKA and basophilic kinases (BB2, AA2, AA1, BB3, BB8, BB9, BB1, AA10, AA9). Various clusters of tyrosine kinases are also observed (XX2, XX1, WW3 and YY2, YY9, WW1, WW9, WW8). Focusing on likely kinase-substrate pairs, the Pearson correlation coefficient between KK2 and AKT2 was 0.88 whilst acidic substrate DD9 showed very good correlation with casein kinase 2 (CSNK2A2) with a Pearson correlation coefficient of 0.97(Figure 4.13).

In general, broader profiles, spanning more fractions, are often observed for the kinase activity profiles compared to the protein abundance profile. This is likely to be on account of the greater sensitivity of the kinase activity compared to the spectral counting method used for quantitation of the protein abundance. This unfortunately reduces the accuracy of the correlation for proteins with low spectral counts. For example, the kinase CDK7 shows good correlation with the peptide substrate PP9 in fractions 14-19 (Figure 4.13). However, in the neighbouring fractions, whilst kinase activity is still detected (fractions 11-14 and 19 to 32), very bad correlation is observed with the kinase spectral count abundance on account of the limited sensitivity of spectral counting.

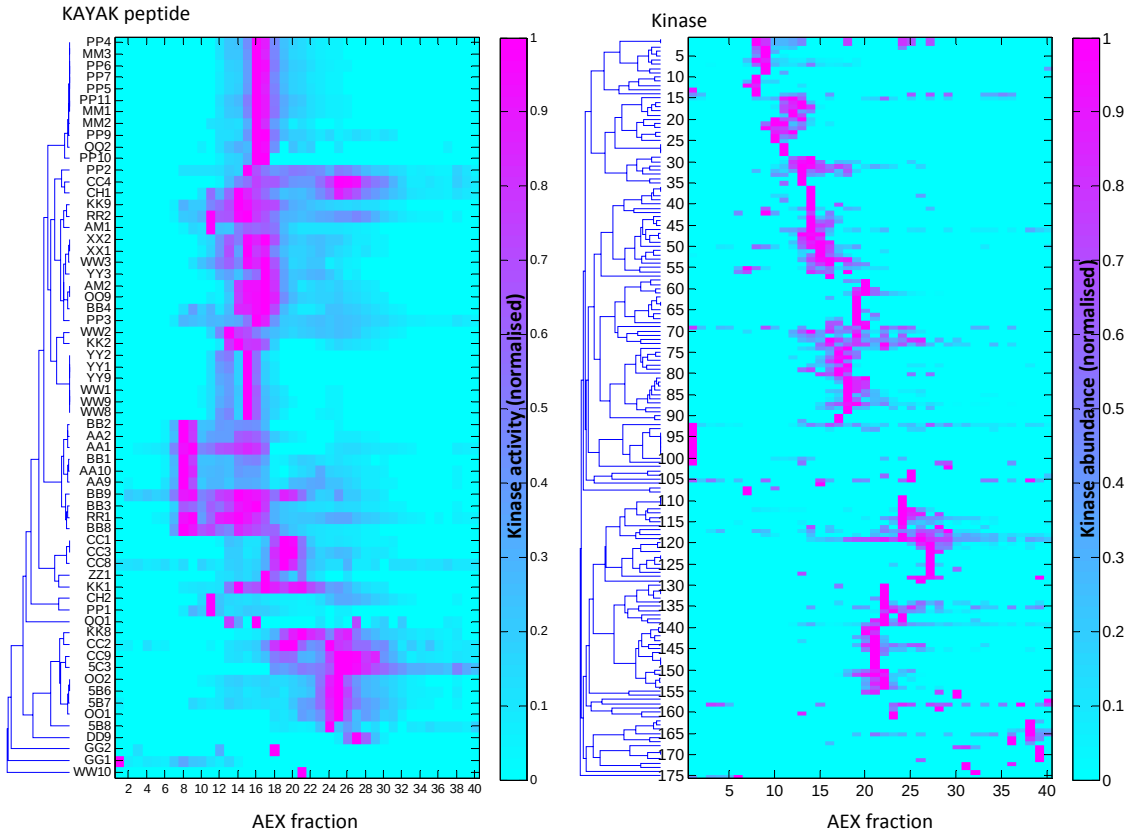


Figure 4.11 Heat map of (a) kinase activities and (b) kinase abundance from HeLa cell nocodazole arrested cell lysate AEX fractions. Kinase activities determined by K60 substrate set on each AEX fraction for (a). Kinase abundance determined by shotgun proteomics and LC-MSMS analysis for (b). List of proteins labelled 1-175 (b) is detailed in Appendix M. Analysis by Pearson correlation hierarchical clustering was used to group similar responders together. Each row represents the phosphorylation rate of a different peptide (a) or kinase abundance (b) normalised to the maximum value across each row.

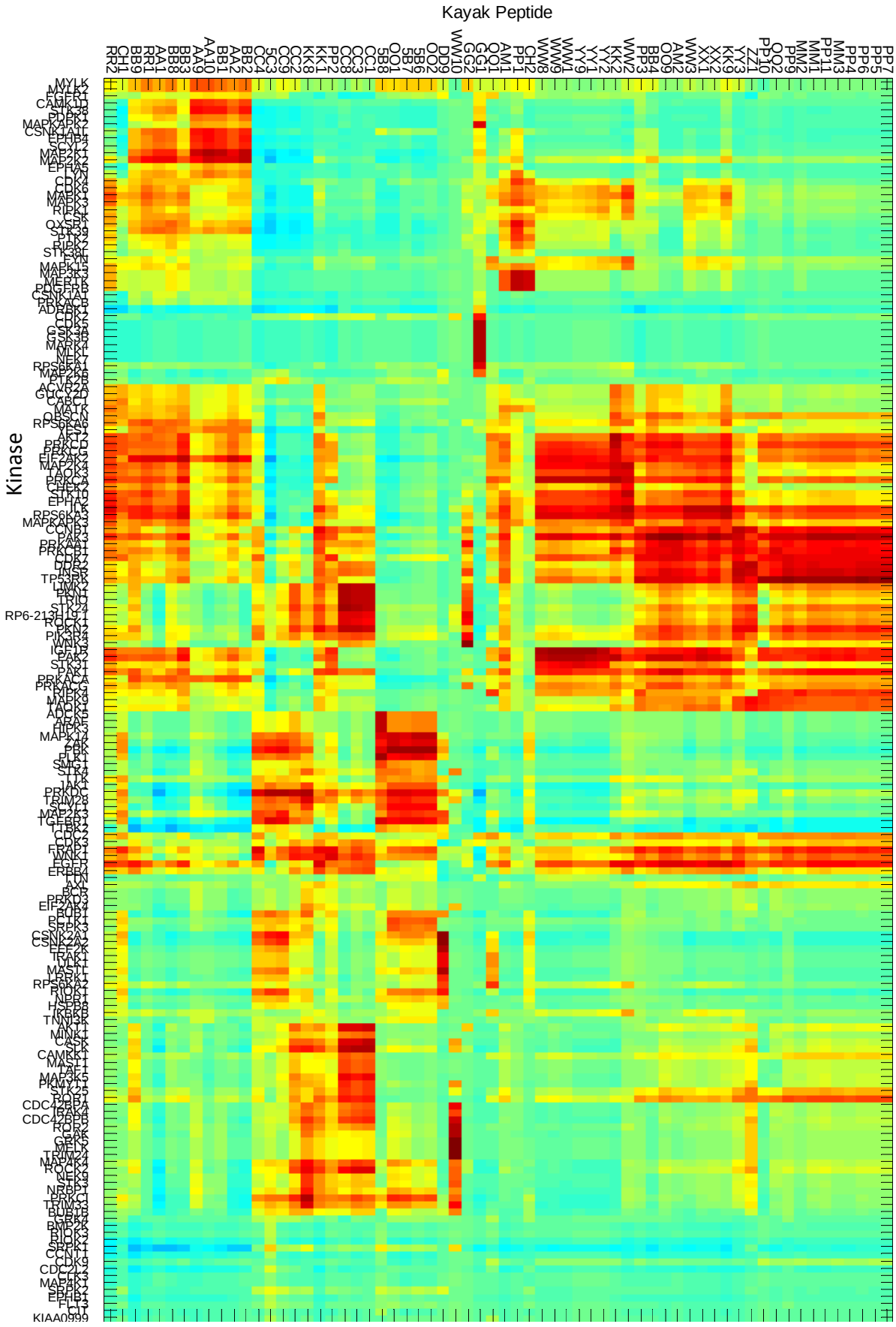


Figure 4.12 In-gel KAYAK correlation profiling using AEX for HeLa cell nocodazole arrested cell lysate. Heat map of Pearson correlation coefficients between the Kinase activity profile (KAYAK) and Kinase abundance profile (spectral count) across the 40 AEX fractions.

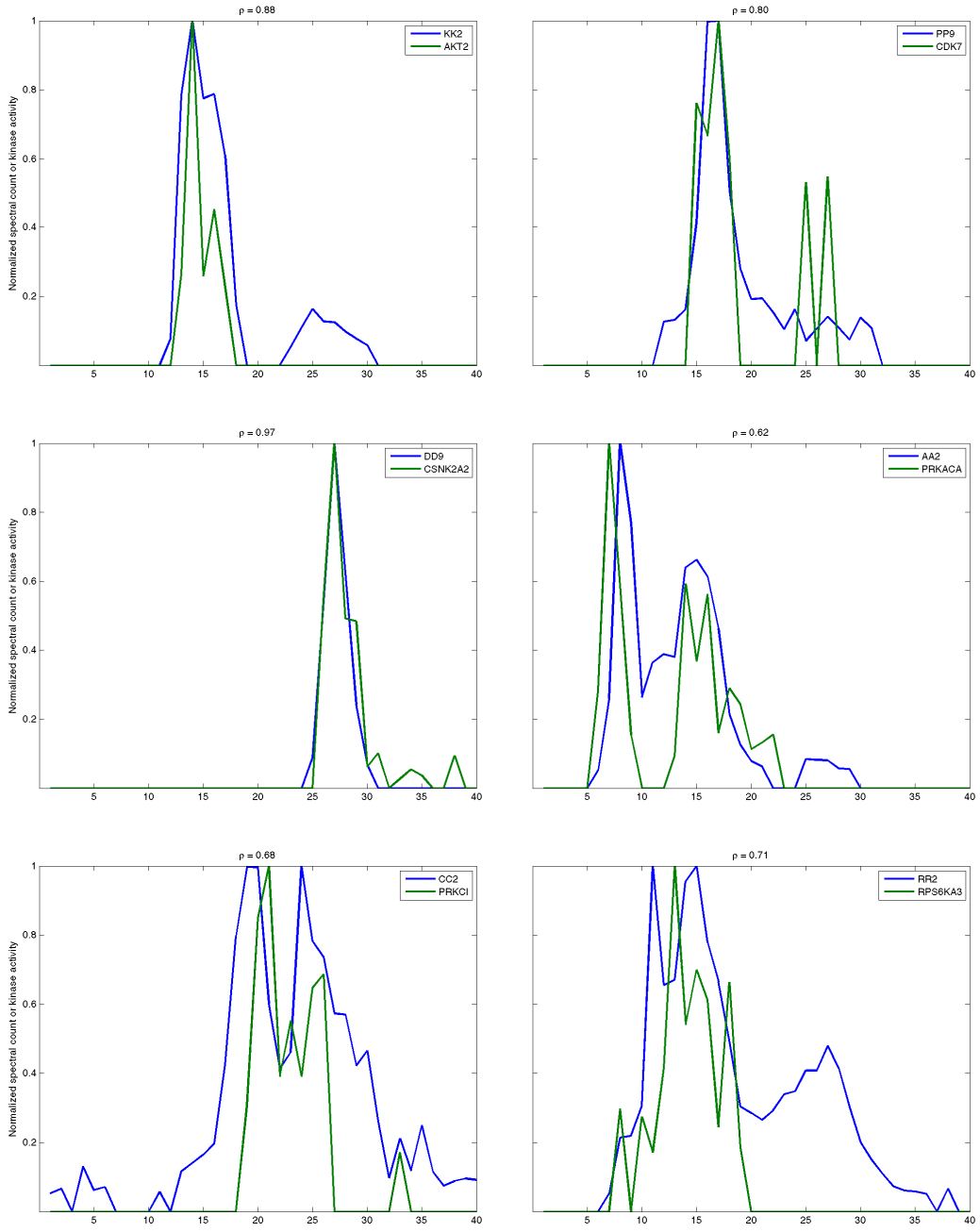


Figure 4.13 Correlation profile plots of selected known kinase-substrate pairs. The normalised spectral count and KAYAK activities are plotted for each AEX fraction. The Pearson correlation coefficient (ρ) for each kinase-substrate pair is given above each plot.

4.3.4 Correlation profiling to identify kinases responsible for the phosphorylation of proteins during mitosis

In the previous analysis it was shown that AEX KAYAK correlation profiling was able to identify known kinase-substrate pairs. The following analysis describes the application of KAYAK correlation profiling to determine the identity of the kinases responsible for the phosphorylation of substrates where the responsible kinase is not known.

It is of interest to characterise the cellular machinery which can lead to genomic instabilities as a way to better understand malignant disease mechanisms. The genes involved are not only of prognostic value, but also serve as cancer therapeutic targets. Many cancers involve altered regulation of the cell cycle and therefore greater understanding of cell cycle regulation and progression is of great value.

4.3.4.1 Kinase activity assay on peptides from mitotic phosphorylation study

From a study of mitotic phosphorylation by Dr. Noah Dephoure in the Gygi laboratory[176], 10 peptides with interesting motifs and mitotic phosphorylation activity were selected (Table 4.1). It should be noted that the cognate kinase for each selected peptide was unknown. To determine whether differences in phosphorylation of these peptides occur, KAYAK was performed on asynchronous (AS) and nocodazole arrested (G2/M) HeLa cell lysate using the selected 10 peptides at 5 µM substrate concentration. Significant differences in kinase activities were detected between AS and G2/M HeLa lysate against 3 peptides: 5B6, 5B7 and 5B8. Very little kinase activity was detected for the other 7 peptides. The assay was repeated using a higher concentration (50 µM) of the seven peptides and resulted in significant differences in activity being detected between G2/M versus AS for peptides 5B4, 5B10 and 5B11 (Figure 4.14).

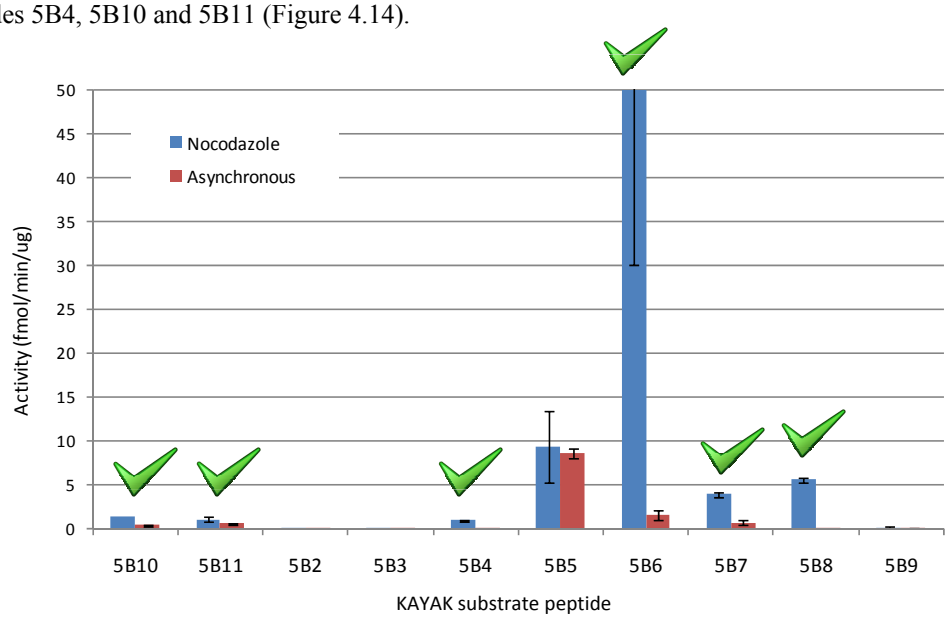


Figure 4.14 Kinase activities for the 10 peptides in nocodazole arrested and asynchronous HeLa cell lysate. 5B6, 5B7 and 5B8 are at 5 µM, 5B4, 5B10 and 5B11 are at 50 µM. 5B2, 5B3 and 5B9 do not have significant activities and were not considered further. Whilst 5B5 has significant activity there was no significant different in G2/M versus AS and therefore was not included in further experiments.

4.3.4.2 Kinase activity-abundance correlation profiling on mitotic peptides

To identify the kinases responsible for the mitotic peptides of interest (5B6, 5B7, 5B8, 5B4, 5B10 and 5B11), a kinase activity assay correlation profiling experiment was performed. Separation was performed using AEX, and HeLa cell nocodazole arrested lysate was divided into 40 fractions, each then being analysed using KAYAK (to determine the kinase activity) and shotgun proteomics (to determine the kinases present in that fraction). By correlating the kinase activity with the kinase abundance profile, putative kinases responsible for the phosphorylation of each particular peptide were obtained.

For each of the 40 fractions, KAYAK was performed using a K8 substrate set which comprised the six mitotic peptides which were differentially phosphorylated in asynchronous versus nocodazole arrested HeLa cells (Table 4.1) and two peptide controls (B11 and C2) whose extent of phosphorylation is known to increase at G2/M . The mitotic peptides 5B6, 5B7 and 5B8 were at 5 µM whilst 5B4, 5B10 and 5B11 were at 50 µM substrate concentration, and the K8 kinase activities were measured across the 40 fractions to obtain kinase activity profiles (Figure 4.15). Of the K8 peptides: 5B6, 5B7 and 5B8 had similar normalised profiles (labelled peak 3 in Figure 4.15), 5B4, 5B10, B11 and C2 (labelled peak 2 in Figure 4.15) also had similar normalised profiles, whilst 5B11 overlapped slightly with peak 2 but was slightly offset (labelled peak 1 in Figure 4.15). This suggests that a single kinase might be responsible for the phosphorylation of 5B6, 5B7 and 5B8 (peak 3), another kinase for peak 2 (5B4, 5B10, B11 and C2) and another kinase for 5B11 (peak 1).

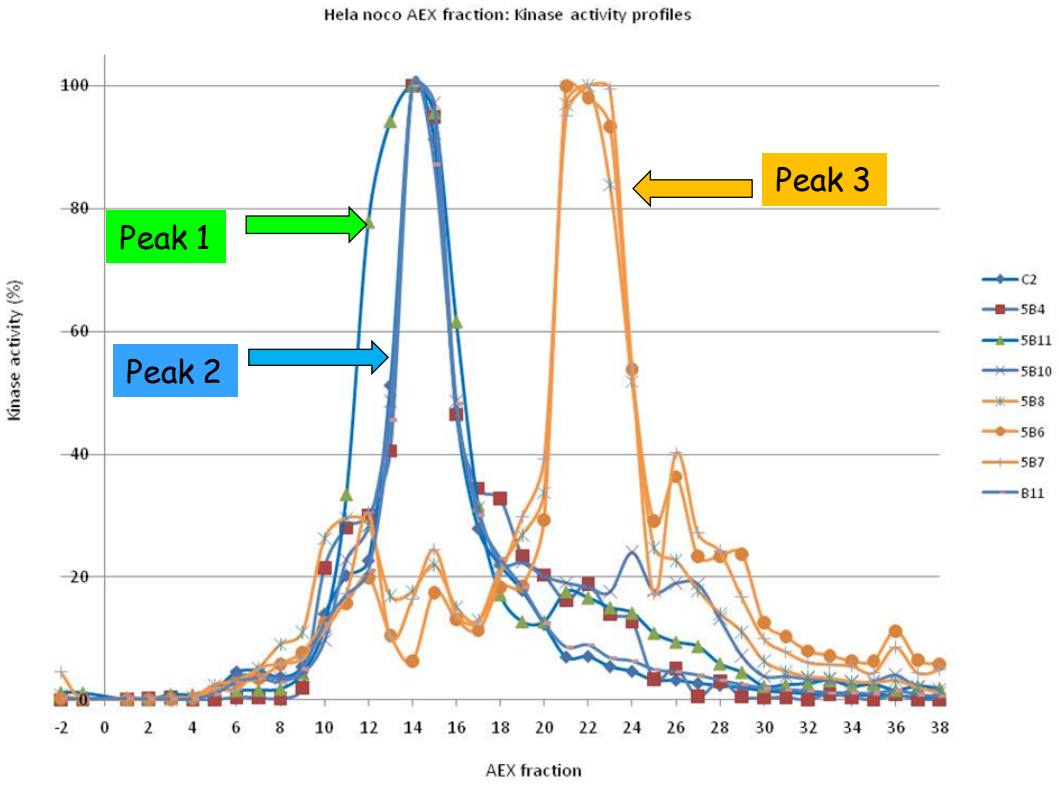
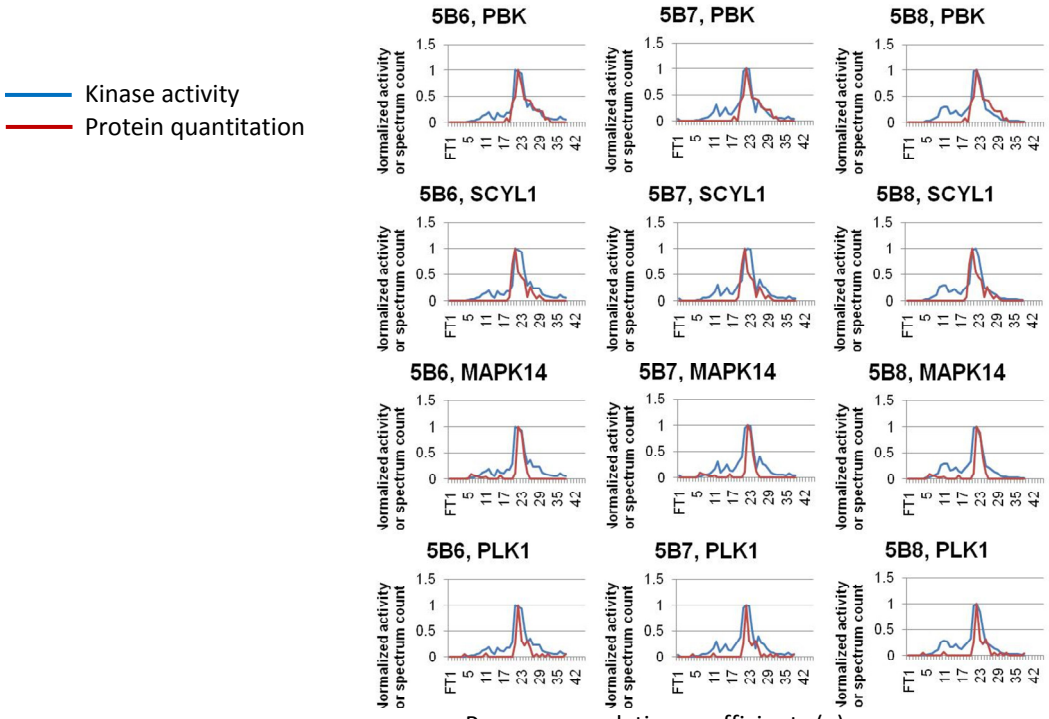


Figure 4.15 KAYAK correlation profiling on HeLa nocodazole arrested cell lysate. KAYAK profiles for K8 KAYAK substrate peptides across 40 fractions of AEX fractionated HeLa nocodazole arrested cell lysate. Kinase activities for each of the peptides are normalised to the fraction in which the greatest kinase activity was observed.

Shotgun sequencing by LC-MS/MS of each of the 40 fractions resulted in the identification of 7150 proteins (peptide FDR 0.41%, protein FDR 10.2%). Spectral counting of proteins annotated with gene symbols resulted in the quantification of 5626, 177 of which were kinases. To correlate the protein profile with the kinase activity, the Pearson correlation coefficient was calculated for each protein for each kinase activity profile. For the peptides 5B6, 5B7 and 5B8, the Pearson correlation coefficients of the top four proteins ranged from 0.75-0.90 and included the proteins PBK, SCYL1, MAPK14 and PLK1 (Figure 4.16).



Pearson correlation coefficients (ρ)

	5B6	5B7	5B8
PBK	0.90	0.87	0.81
SCYL1	0.85	0.84	0.80
MAPK14	0.80	0.79	0.78
PLK1	0.79	0.75	0.78

Figure 4.16 Correlation profiles of kinase activity for 5B6, 5B7 and 5B8 and top 4 protein kinase profile hits. Pearson correlation coefficients for kinase activity and protein profiles for 5B6, 5B7 and 5B8 with the top 4 kinase hits.

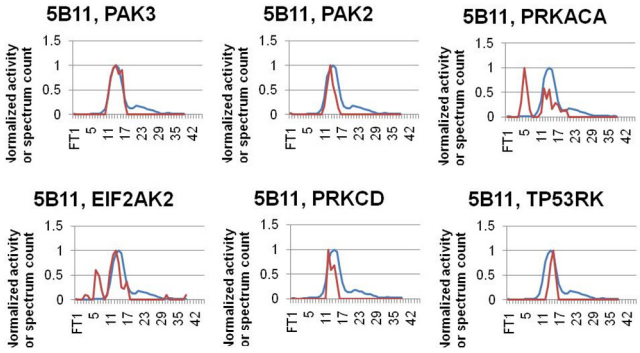


Figure 4.17 Correlation profiles of kinase activity for 5B11 and top 6 protein kinase profile hits.

The top ranked putative kinase for peptide 5B11 was PAK3 with a Pearson correlation coefficient of 0.97 (Figure 4.17). The coefficients for the next top possible hits (PAK2, PRKACA, EIF2AK2, PRKCD and TP53RK) were much lower. PAK3 was the first ranked kinase for 5B11 and was the third ranked protein out of the 5626 proteins. For peak 2 (Figure 4.15), which comprises the peptides 5B4 and 5B10 as well as the two peptides which were added as standards (B11 and C2), the top hit was TP53RK (Figure 4.18).

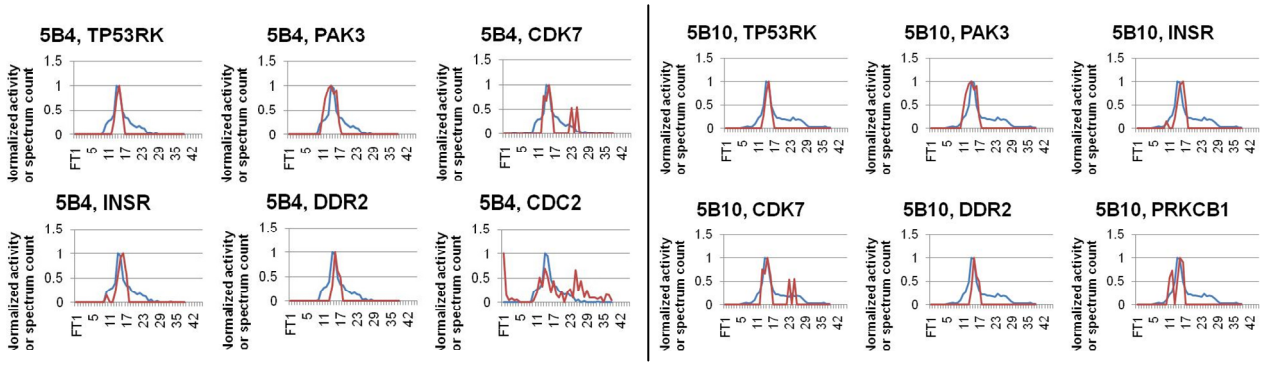


Figure 4.18 Correlation profiles of kinase activity for 5B4 and 5B10 and top 6 protein kinase profile hits for each.

One potential source of error with the current correlation-based approach is that the quantitation using spectral counting may have some level of inaccuracy, especially for low abundance kinases with only one or two spectral counts. Improved protein profiles might be obtained by using raw ion currents (label-free), aqua peptides to quantitate each of the kinases, or stable isotope labelling of consecutive fractions. Another approach to remedy this situation would be an enrichment of kinases using kinobeads[321] or Proteominer beads (Biorad) to be able to identify more kinases. Alternatively, a two dimensional approach could be used whereby the proteins are fractioned twice, for example the first dimension on a gel filtration column and the second dimension using AEX. All these approaches will be investigated in the future. Regardless of the possible inaccuracies in the quantification of the substrate pairs, interesting biological insights have been obtained.

To improve the accuracy of the protein profiles (kinase abundance) and to determine how variable the spectral counting profiles were, fractions 10-20 were shot on the mass spectrometer 3 times. Although there were some differences in the ranking order and correlation coefficients, illustrating the variability in the spectral counting protein quantitation, the top ranked proteins were similar to those found previously.

Since there were several high ranking kinases for the putative kinase-substrate pairs, it was necessary to perform validation experiments to confirm which kinase was responsible.

4.3.4.3 Validation experiments - Kinase assays using recombinant kinases

The standard peptides B11 and C2 are known to be phosphorylated by Cdc2/cyclin B[150]. Since the KAYAK profiles of 5B4 and 5B10 co-elute with the B11 and C2, 5B4 and 5B10 may also be phosphorylated by Cdc2/cyclin B. To validate this hypothesis, KAYAK was performed on the K8 peptides using purified Cdc2/cyclin B kinase (Table 4.2). The standard peptides B11 and C2 were significantly phosphorylated. However, the remaining peptides, including 5B4 and 5B10, fell below the limit of accurate detection (50-100 fmol), indicating that Cdc2/cyclin B was not responsible for their phosphorylation.

Table 4.2 Product formed by *in vitro* kinase reaction of Cdc2/cyclinB with K8 substrate mix

peptide	Sequence	Product (fmol)
B11	LKLSPsPSSRPFR	536 ± 39
C2	IPTGTtPQRKPFR	2966 ± 106
5B10	EFAAALtAERIKTfK	31 ± 9
5B11	HFRNVHsEDFENRfK	72 ± 31
5B4	KPSQVSsGQKLGPfK	0
5B6	DDFKLNSsIVEPKfK	0
5B7	IEKLNSsLHFLQQfK	0
5B8	AQHPDYsFGELSRfK	0

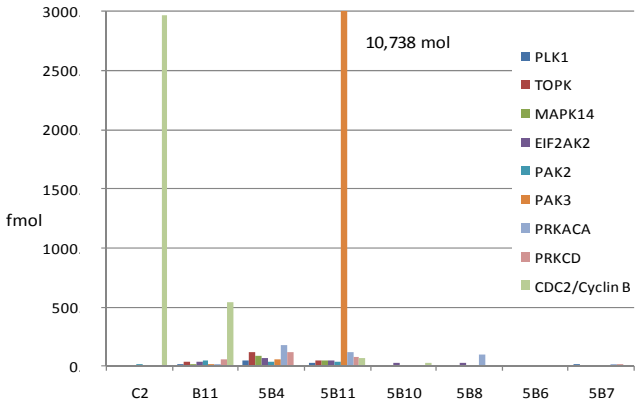


Figure 4.19 Product formed (fmol) for K8 peptides using purified kinases.

Subsequent investigation involved the elucidation of the responsible kinase activity and validation of the results from the correlation profiling experiments. For each putative kinase that was commercially available, KAYAK was performed on the K8 substrate mix (Figure 4.19). Phosphorylation of 5B11 by purified PAK3 was clear, validating the results from the correlation profiling experiment. Apart from the standard peptides (C2 and B11) which were phosphorylated by CDC2/cyclinB, none of the other purified kinases phosphorylated the peptides of interest. There are several possible explanations for this result. Firstly, the purified kinase may not be active, a common problem with commercial kinases. Secondly, kinase activity may require the presence of a cofactor, or only be active when it is present in a complex.

To address the first point (active kinase): PLK1 and PBK were obtained from various manufacturers (CST, Millipore, SignalChem) and no significant phosphorylation of the peptides of interest (5B6, 5B7, 5B8) was observed using any of these kinases. To address the second point (presence of cofactors), KAYAK assays were performed on the purified kinases in the presence and absence of cell lysate. The presence of cell lysate did not appear to increase the activity of the kinases PLK1 and PBK for the peptides 5B6, 5B7 or 5B8. However, there was evidence that EIF2AK2 in the presence of nocodazole arrested lysate led to some increased phosphorylation of 5B4 and 5B10 compared to EIF2AK2 or lysate alone. This may indicate that a cofactor needs to be present for EIF2AK2 to be active, and that EIF2AK2 could be the responsible kinase for the phosphorylation of 5B4 and 5B10. This is discussed in further detail in Appendix N. Finally, to address both issues simultaneously (to ensure that the kinase was active and that its binding partners were present) the kinases were overexpressed (using Gateway cloning) and knocked down (using siRNA) in human cells.

4.3.4.4 Validation experiments - kinase assays using overexpressed kinases

To overcome the problems associated with using purified commercial kinases, and to ensure the kinases were active and present with the relevant cofactors and binding partners, various kinases were overexpressed. Overexpression was performed in HEK 293 T cells using Gateway cloning. KAYAK was performed on each of the overexpressed kinase lysates for both K9 and K60 substrates. The kinase activities for 5B6, 5B7, 5B8 were significantly upregulated by 2- to 3-fold in the lysate with overexpressed PLK1 (Figure 4.20a and b). Along with PLK1, TOPK and MAPK14 were the the other two top putative kinases. However, upregulation of kinase actvity was not detected for either overexpressed TOPK (PBK) or MAPK14. Whilst the overexpression of PLK1 leads to increased phosphorylation of the the substrates 5B6, 5B7 and 5B8, it does not confirm whether PLK1 is a direct or indirect agent of phosphorylation. Overexpression of PLK1 may be responsible for the activation of another kinase which is then responsible for the increased phosphorylation of 5B6, 5B7 and 5B8.

Unfortunately, constructs for all the kinases of interest were not available or their transfection was not successful, and siRNA knockdown of several kinases was investigated including TP53RK, CDC2, PAK3, MAPK14 and TOPK (PBK). Unfortunately, siRNA PLK1 is lethal to the cell and therefore this kinase cannot be knocked down. An alternative would be to immunodeplete PLK1, and this will be performed in the future. The results of the knock down experiments were inconclusive and are discussed in Appendix O.

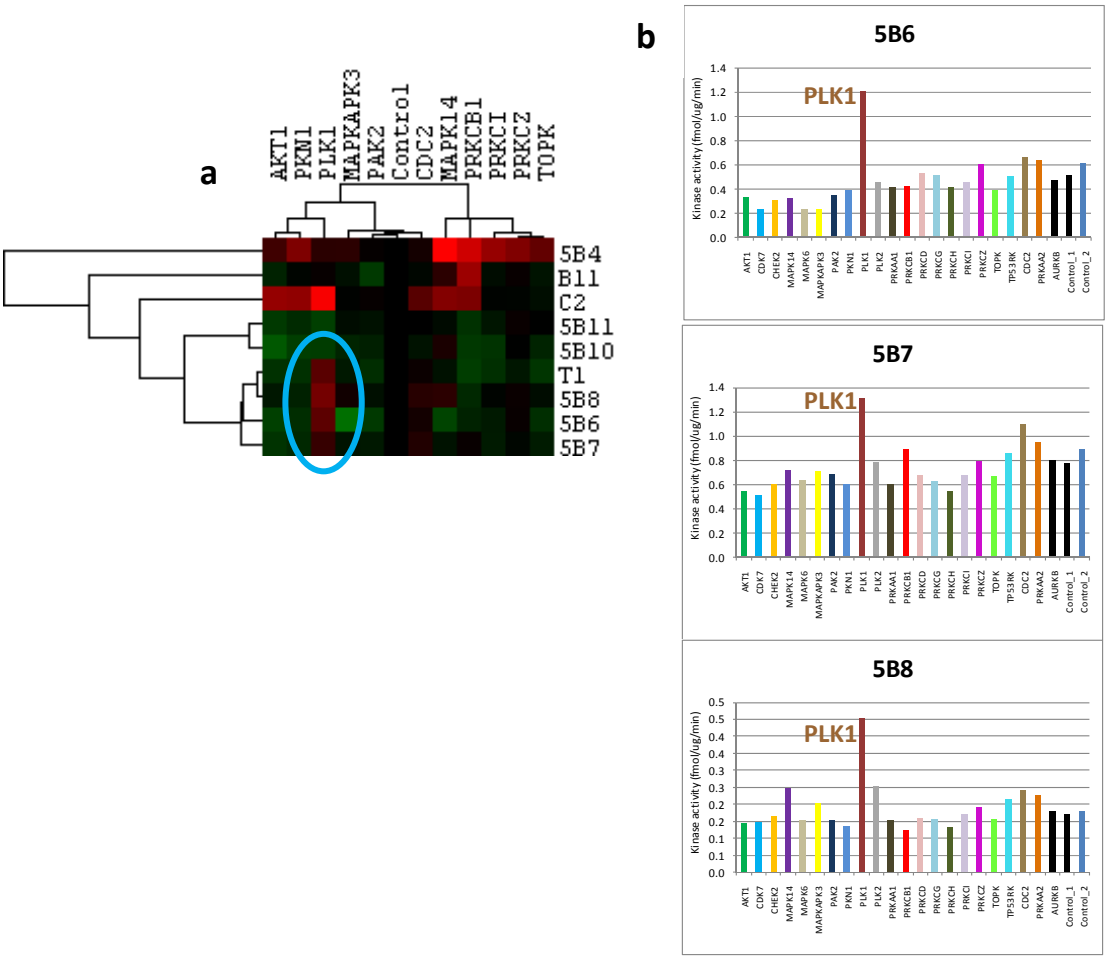


Figure 4.20 Overexpression of kinases in HEK 293T cells using Gateway cloning. (a) Heat map of kinase activities for K9 substrates with overexpressed kinases. (b) Kinase activity graphs for 5B6, 5B7, and 5B8 for each of the overexpressed kinases.

From the correlation profiling data of the mitotic peptides, PAK3 is responsible for the *in vitro* phosphorylation of 5B11 and PLK1 appears to be responsible for the phosphorylation of 5B6, 5B7 and 5B8. Further validation experiments need to be carried out to determine which is the responsible kinase phosphorylating 5B4 and 5B10. TP53RK would appear to be the top hit. However, the recombinant kinase is not available commercially and the overexpression of TP53RK was not successful.

For the putative kinase-substrate pairs (PAK3-5B11), (PLK1-5B6, PLK1-5B7, PLK1-5B8), the correlation profiles of proteins that are known to interact with the kinases of interest were considered. Since the presence of particular binding partners is sometimes required for the kinase activation, the presence of such partners could help indicate whether the kinase is active or not. For example, it is known that Cdc2/cyclinB phosphorylates peptides B11 and C2, but Cdc2 alone is not able to phosphorylate B11 and C2. The presence of interacting partners of PAK3 and PLK1 are discussed in Appendix P.

4.3.4.5 Biological importance

The KAYAK correlation profiling platform allowed for the identification of four potential kinase-substrate pairs (Table 4.3). These findings are novel associations of the kinase PAK 3 to substrate 5B11 (derived from protein ADNP) and the kinase PLK1 to substrates 5B6, 5B7 and 5B8 which are derived from the proteins NFKB2, CAMSAP1L1 and PB1 respectively.

Table 4.3 Putative kinases for peptide substrates 5B10, 5B11, 5B4, 5B6, 5B7 and 5B8 and a summary of the experiments providing the evidence.

Peptide	Sequence	From protein	Top Putative kinase	Evidence			
				Recombinant kinase	recombinant kinase + lysate vs. lysate alone	Over-expressed kinase	siRNA knockdown
5B10	EFAAALtAERIKTfK	EZH2	TP53RK/ EIF2AK2	no	Some evidence in favour of EIF2AK2	N/A	no
5B11	HFRNVHsEDFENRfK	ADNP	PAK3	yes	yes	N/A	no
5B4	KPSQVSsGQKLGPfK	BMS1L	TP53RK/ EIF2AK2	no	Some evidence in favour of EIF2AK2	N/A	no
5B6	DDFKLNSsIVEPKfK	NFKB2	PLK1	no	no	yes	N/A
5B7	IEKLNSsLHFLQQfK	CAMSAP1L1	PLK1	no	no	yes	N/A
5B8	AQHPDYsFGELSRfK	PB1	PLK1	no	no	yes	N/A

4.3.4.5.1 PAK3

The sequence of the 5B11 substrate peptide found in ADNP (Activity-dependent neuroprotector homeobox) has an arginine at the -4 position and a phenylalanine (a hydrophobic residue) at the +3 position. This is consistent with the reported residue preferences reported by Rennefahrt *et al.*[322] and Zhu *et al.*[323], and thus it is likely that PAK 3 is responsible for phosphorylation of 5B11. ADNP is a chromatin binding protein that plays a role in cell-cell adhesion and the regulation of gene expression during embryogenesis. ADNP is a neuroprotective factor and can either stimulate or inhibit the growth of particular tumour cells. ADNP contains 9 zinc finger domains and one homeobox and it is a putative transcription factor. ADNP is known to be critical for brain development.

PAK3 is a member of the p21 activated kinase (PAK) family and a member of the MAPKKK orthofamily. There are 6 known PAK isoforms that are split into 2 groups: Group I PAKs (PAK 1-3) and Group II PAKs (PAK 4-6)[324]. The 2 groups are split based on their sequence and regulatory mechanism similarities. PAKs are targets of the Rho family GTPases RAC and CDC42, and Group I PAKs are significantly upregulated on binding to GTP-bound RAC or CDC42. Group II PAKs are known to bind RAC and CDC42 but do not appear to be activated by the binding[322]. The roles of PAKs are not known exactly. PAKs appear to be involved in multiple signalling pathways in a subcellular location-dependent manner involving ERK2 and JNK[325]. As well as having general roles in cellular processes including actin cytoskeletal reorganisation, it is likely that PAKs also play a role in specialised activities such as neuronal cell differentiation and motility[325]. The proposed phosphorylation of ADNP by PAK3 is interesting since ADNP is known to be critical for brain and neuron development and PAKs are thought to play a role in neuronal cells and this work offers a link between the two which has previously not been reported. The role of PAK3 could be important for the activation of ADNP which may contribute to its key role in brain development.

4.3.4.5.2 PLK1

PLK1, polo-like kinase, is a serine-threonine kinase and is involved in the G2/M transition in the mitotic cell cycle. PLK1 is highly upregulated in G2/M, along with Aurora and cyclin-dependent kinases[176]. There are 4 known polo-like kinases in mammalian cells: PLK1, PLK2, PLK3 and PLK4[326]. All the PLKs have a similar structure comprising an N-terminal S/T kinase domain and a C-terminal polo-box domain (PBD) that comprises polo box 1 and polo box 2 with a linker in between. PLK1 is known to be involved in many events in mitosis, particularly during M phase. PLK1 phosphorylates Cdc25C and cyclin B1, resulting in their translocation to the nucleus. PLK1 is known to phosphorylate the cohesion subunit SCC1 which allows cohesion to be displaced from the chromosomes to centromeres. PLK1 phosphorylates various APC (anaphase promoting complex) subunits, where APC activation is required for mitotic exit[327].

The PLK1 motif, according to the PhosphoMotif Finder Human Protein Reference Database (HPRD)[311], is [D/E]X[pS/pT][I/L/V/M]X[D/E]. This is very similar to the sequence for 5B8 (AQHP**D**Ys**F**G**E**LSRfK) which has a phenylalanine (F) residue instead of an I/L/V/M residue. Phenylalanine is hydrophobic, as are isoleucine (I), leucine (L), valine (V) and methionine (M), and therefore it is likely that PLK1 would phosphorylate this peptide. The peptide sequences for 5B6 (DDFKLNSsIVEPKfK) and for 5B7 (IEKLNSsLHFLQQfK) have very similar motifs, both with KLNSs. However, they are not at all similar to that reported for PLK1, and this data suggests that this could be a new motif for PLK1. In general, PLK1 substrates do contain consensus phosphorylation motifs[328-329], which is consistent with the identification of two motif types in this study.

Peptide 5B6 is derived from NFKB2 (NF–κB: nuclear factor kappa-light-chain-enhancer of activated B cells) which is a very important protein involved in DNA transcription. PLK1 has not been associated with the phosphorylation of NF–κB, therefore the identified link between kinase and transcription factor is a novel and extremely important finding in the control and regulation of mitotic progression.

Protein kinases are predominately responsible for the regulation of cell cycle progression. However, the transcriptional control of cell cycle regulators during mitosis is not well understood. It is therefore important to determine the links between kinases and transcription factors to understand regulation of the cell cycle fully, and also its misregulation which occurs in many cancers[330-331]. In cancers, many cell cycle kinases are often erroneously upregulated and investigation of the interplay between kinases and transcription factors could provide insight into the link between kinase upregulation and cell proliferation through alterations in transcriptional programmes.

Whilst many PLK1 mitotic substrates are known[332-337], to date there has been only one report of a direct link between PLK1 and transcription factors involved in gene expression. It was found that PLK1 directly binds and phosphorylates the mammalian transcription factor Forkhead Box M1 (FOXM1)[338-340] and that this phosphorylation regulates a transcription programme that is essential for

mitotic progression[341]. PLK1 comprises a polo-box domain (PBD) that is responsible for the subcellular localization of PLK1 during mitosis and has been proposed to target the kinase to its substrates[342]. The optimal motif for the PLK1 PBD binding was found to be S-[pT/pS]-(P/X)[343]. Proline (P) in the (pT/pS) +1 position is helpful but not critical for binding. Proline directed mitotic kinases which phosphorylate SP and TP sites can generate PBD binding sites where there is a serine in the -1 position. The PBD in PLK1 is similar to the SH2 domain in the Src family of kinases in that the PBD binds to the kinase domain and inhibits the kinase activity in the basal state. The binding of phosphorylated substrates to the PBD of PLK1 results in it becoming active. Many of the PBD sites for PLK binding are initially phosphorylated by the Cdc2/Cyclin B complex[341].

Examining the sequences of CAMSAP1L1, PB1 and NFKB2 (the proteins from which 5B6, 5B7 and 5B8 are derived), all three proteins have several S-[T/S]-(P/X) motifs (Table 4.4 and Appendix Q). To determine whether there are S-[pT/pS]-(P/X) motifs where the threonine or serine is phosphorylated, data from a phosphoproteomic study of HeLa cells in G2/M was searched[176]. From this data, 3 different S-pS-X sites were found in CAMSAP1L1, one of which is the sequence where the PLK1 kinase is thought to phosphorylate. Two different sites in PB1 were found with S-pS/pT-P. Apart from the site on PLK1 LNSpS, which is the peptide PLK1 is thought to phosphorylate, no other PBD binding motifs were observed in this dataset. The sequences for both 5B6 and 5B7 include a leucine at the -3 position. Similarly, the site on FOXM1 (a transcription factor) which is phosphorylated by PLK1 also has a leucine at the -3 position at the Ser 724 phosphorylation site (IL**L**DIpSFPGL).

Table 4.4 Peptide sequences containing possible PBD binding motifs in CAMSAP1L1, PB1 and NFKB2. Possible PBD binding sites are shown in red. # indicates phosphorylated residue. Data obtained by Dr. Dephoure[176].

Protein Name	site	Ascore	Peptide
CAMSAP1L1	415	5	RSS#SMSYVDGFIGTWPK
CAMSAP1L1	901	17	LNSS#LHFLQQEMQR
CAMSAP1L1	974	7	QAGLSS#AIAPFSSDSPRPTHPSPQSS#NR
PB1	28	6	AT#SPSSSVSGDFDDGHHS#VST#PGPSR
PB1	178	6	GEADDEDDDEDGQDNQGTVTEGSS#PAYLK
NFKB2	23	18	LNSS#IVEPK

The PBD inhibits the kinase activity of PLK1 and therefore if the PBD is not bound to a substrate, such as is the case with purified recombinant PLK1, no activity would be expected. This explains why the purified PLK1 from the commercial sources (Millipore, CST, SignalChem) did not phosphorylate the substrate peptides 5B6, 5B7 and 5B8. However, a peptide known as PoloBoxtide 8T with the core sequence MGSpTPL binds to the PBD and activates the kinase[343]. Adding this into the KAYAK reaction should activate PLK1 and allow the substrate peptides to be phosphorylated by recombinant PLK1. This experiment will be performed in the future. There is also a chemical inhibitor of PLK1 (non-ATP competitive) which will be used to inhibit PLK1 activity[344] for future validation studies.

4.3.4.5.3 PBK

The PBK (PDZ-binding kinase) gene/protein was discovered following a two-hybrid screen[345]. PBK is phosphorylated at mitosis in HeLa cells and this phosphorylation event is critical for its kinase activity. CDC2/cyclin B phosphorylates PBK in vitro. However, *in vitro* phosphorylation of PBK by CDC2/Cyclin B does not appear to activate PBK[345]. Presumably, other posttranslational modifications or the presence of a cofactor are required for PBK activation. PBK binds to the PDZ2 domain in the human homologue of the *Drosophila* Discs-large tumour suppressor protein (hDLG). PBK is hypothesised to have some role in DNA damage sensing and repair[346-347]. PBK also binds to TP53 and is involved in histone phosphorylation[348]. PBK is found to be most abundant in the placenta and is upregulated in various malignancies, particularly haematological ones. There is no published PBK motif for substrate phosphorylation and little is known about its role during mitosis, or signalling pathways in which it has a role.

4.3.4.6 Future validation studies

These initial studies predicting kinase-substrate pairs are a proof-of-concept that, by starting with a 'hit' from a phosphoproteomic experiment[176], KAYAK correlation profiling can identify putative kinases responsible for phosphorylation of a particular substrate peptide.

Given the outcome of these profiling experiments discussed above, further validation studies will be performed using immunodepletion and chemical inhibition of the kinases of interest. Future studies will focus on determining whether the proposed substrate-kinase pairs are direct or indirect substrates. Shokat and co-workers have developed various mutant kinases which can be specifically and rapidly inhibited by the pyrimidine-based inhibitor 1-NM-PP1[349]. The Shokat team has also developed mutant kinases that can only bind bulky ATP. Bulky ATP is an analogue of ATP with bulky substituents synthesised in the N-6 position and the mutant kinases are created by introducing a space-creating gatekeeper mutation into the kinase ATP binding pocket. This allows the mutant kinase to bind bulky ATP which cannot be utilised by wild type kinases. By comparing the mutant PLK1 cell lines, both with the inhibitor and with a bulky ATP and controls, it will be possible to determine whether the substrates are direct or indirect.

Further studies will also involve investigating binding partners of PLK1 and PAK3 using immunoprecipitate mass spectrometric analysis to determine whether the proteins that are pulled down are those that have similar profiles to those obtained using the anion exchange chromatographic profile.

4.4 Conclusions

The multiplexed kinase assay with a 60 peptide substrate set (K60) was used to analyse a wide variety of tissues from different organisms; including human saliva, human cell lines, tissues from mouse, *Drosophila* embryos and yeast. Significant kinase activity was observed in human saliva. The source of kinase activity in saliva is not known but proposed to be lysed cells from the lining of the mouth. Although the kinase substrates were designed for human kinases, many of the substrates are also phosphorylated by kinases in other organisms. This illustrates that the kinase substrate motifs are conserved between species and demonstrates the wide applicability of KAYAK as a cross species kinase assay.

Large differences in kinase activity were observed in the different mouse tissues, and in general there was good correlation between the kinase abundance (spectral count) profile in the different tissues and the corresponding kinase activity of the putative kinase-substrate pairs. This is the first report of a kinase activity/kinase abundance correlation profiling comparison between different tissues.

Proof-of-concept work was performed for an in-gel kinase activity assay and kinase activity/kinase abundance correlation profiling using blue native polyacrylamide gel electrophoresis. However, low resolution of the gels limits the applicability of this technique for identification of kinase-substrate pairs. Kinase activity/kinase abundance correlation profiling was performed using anion exchange chromatography to identify the kinase responsible for the phosphorylation of substrates during mitosis and, as a result, novel kinase-substrate pairs have been proposed. This illustrates the power of the technique to identify kinases responsible for unknown substrate phosphorylation events.

Correlation profile identification of the kinase responsible for a specific phosphorylation event was also performed. This study showed that the sample complexity was a challenge and several candidate kinases were selected for validation experiments. To narrow down the number of candidate peptides, two dimensional chromatographic separation could be carried out; for example, gel filtration chromatography followed by anion exchange chromatography could be performed.

KAYAK has huge potential in diagnostics, drug development and testing, and research into cause of disease and signalling pathway disruption. There are several different ways the KAYAK assay could be further developed. This includes optimisation for the testing of drugs such as kinase inhibitors, or alternatively, it could be envisaged as having a role in hospitals for disease diagnosis. Furthermore, various KAYAK kits with peptides specific for particular diseases could be developed. A general peptide kit, along the lines of the K60, could be used for ‘broad diagnosis’ (e.g. healthy/diseased). More specialised peptide kits could aid in diagnosis of specific pathway dysfunctions, including peptides whose phosphorylation state can differentiate subtle changes in signalling cascades. KAYAK may also be used to process biopsy samples, such as tumours. This technique would be particularly useful in cases where both healthy and cancerous tissue can be obtained from the same patient, renal cell carcinoma or skin cancers for example, allowing for a patient to act as their own

control. Beyond cancers, there are many other diseases that could benefit from analysis by KAYAK. For many diseases KAYAK could help characterise the signalling pathways which are altered and possibly lead to the development of kinase-related diagnostic tools. The detection of kinase activity in saliva and also in other body fluids (e.g. tear and serum) opens up another large diagnostic field.

Another possible enhancement to KAYAK would be to extend the K60 kit to include a very large number of peptides, perhaps to the order of thousands, for research purposes. Although many kinases across different species recognise similar motifs, it would be of value to provide KAYAK peptide kits specially tailored for individual species. In terms of kinase-motif evolution it would be interesting to compare the KAYAK peptides of a much greater range of organisms from different phyla. It would be particularly interesting to study the KAYAK activity present in tissues of the nematode *C. elegans,* due to the relatively conserved nature of the kinase families between humans and *C. elegans*. Whilst kinase evolution can be investigated using genomics, this does not determine whether the kinase is active or not. KAYAK, on the other hand, can assess the kinase activity and also assess the conservation of the motifs that the kinases recognise.

The kinase activity/kinase abundance correlation profiling method offers a unique way of identifying kinase-substrate pairs. Profiling different tissues allows for determination of the particular location in which a kinase is active. The location of kinases in different tissues and different parts of the cell affects their activation. By performing kinase activity assays and kinase abundance profiling on different cellular preparations (e.g. membrane, endoplasmic reticulum, nucleus, mitochrondria), information and insight about the localisation of kinase-substrate pairs in different portions of the cell can be obtained. By performing kinase activity/kinase abundance profiling in this way kinase-substrate pairs should be easier to identify.

As well as investigating kinase activity, the KAYAK strategy and the kinase activity/kinase abundance correlation profiling approach has the potential to be applied to numerous other enzyme classes. The results presented here, taken together, demonstrate the wide applicability of the KAYAK method for understanding protein function, and highlight it as a tool with great potential for research and in the clinic.

5 Saliva biomarkers for gingivitis

5.1 Introduction

This chapter integrates the work on salivary proteomics as discussed in Chapter 3 with the work on saliva kinase activity profiling as discussed in Chapter 4 in the context of two clinical trials. The first trial comprised a small study where saliva samples were obtained from 10 gingivitis patients and 10 healthy subjects. This study was performed 'in-house' at the Proctor & Gamble site in Rusham Park, Egham. The second trial, carried out at the Eastman Dental Institute (London, UK), involved induction of gingivitis in healthy subjects via abstention of teeth brushing.

Chapter 3 detailed the work investigating the salivary proteome and analysing the data using three different label-free software packages, with the goal of developing a workflow for comparison of the clinical trial saliva samples. However, because none of the label-free software packages for use with the ion intensity data were suitable for analysis of the saliva clinical trial samples, a more simple method, spectral counting, was used. In this case, the number of spectral counts per protein was used as an estimate of protein abundance.

The saliva samples from the clinical trials were analysed by proteomics. The saliva samples from the induced gingivitis clinical trial were also analysed by KAYAK to determine the kinase activity of both the supernatant and the residue. In the second clinical trial (induced gingivitis), both the saliva supernatant and the residue were investigated and searched against both human and bacterial protein databases. It is hypothesised that gingivitis is caused by, or is a result of, large-scale changes in bacterial populations and activity. If true, this suggests that investigating quantitative changes in bacteria protein abundance and activity may provide a useful method for gingivitis biomarker discovery. Despite the potential wealth of clinically-relevant information which may be present in proteomic analysis of saliva residue, very few past studies have investigated either the bacterial or human cellular protein profiles in saliva residue[211].

5.1.1 Gingivitis and periodontitis

Gingivitis and periodontitis are gum diseases with symptoms including bleeding, swollen, and receding gums (Figure 5.1). Gingivitis is a mild form of periodontitis and is reversible, whilst periodontitis is not.[292] Gingivitis lesions can spread to deeper, supporting tissues and lead to destructive periodontitis lesions. However, this progression does not always occur and the factors that affect this progression are not fully understood. The primary cause of periodontitis is the build up of bacterial plaque that is often a result of inadequate oral hygiene. This build up of plaque results in inflammation leading to the degeneration of epithelial cells at the junction of the tooth and gingiva. This inflammatory lesion may become more extensive, leading to the destruction of dentogingival

alveolar crest fibres and connective tissue down the root surface[101]. With the destruction of connective tissue, an irreversible periodontal pocket forms and can eventually lead to tooth loss.

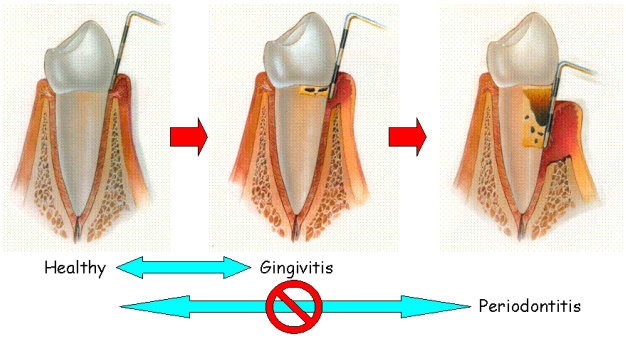


Figure 5.1 Schematic illustrating gingivitis and periodontitis. Gingivitis is considered a milder form of periodontitis and the symptoms can be reversed, unlike periodontitis. Depth of probing (shown in figure as a white and black rod) is used as a clinical indication of the severity of gingivitis/periodontitis. Schematic adapted from Bostrom *et al.*[350]

While the build-up of plaque is the primary cause of gingivitis, various factors, including the composition of the local bacterial flora, affect the overall progression of the disease. Inflammation of the gingiva is thought to start when endotoxin or bacterial lipopolysaccharides induce the release of proinflammatory cytokines by gingival cells[351]. It has been suggested that there are between 6-12 bacterial species which are particularly prevalent in progressive periodontitis[352,353]. Although oral hygiene plays a large role in periodontal disease, other factors play a role. It has been reported that, independent of oral hygiene, around 10% of patients with gingivitis have a high risk of developing destructive, rapid progressing periodontitis and tooth loss; around 80% have susceptibility to periodontitis but progression is slow and normally doesn't result in tooth loss; and around 10% are relatively resistant to destructive periodontitis[354]. These differences are likely a result of the genetic constitution of the patient. The progression of gingivitis to periodontitis is dependent on the interactions between periodontopathic bacteria and the cells of the host immune system. These interactions are mediated by the various cytokines and chemokines secreted by leukocytes and gingival tissues.[355] These molecules are produced at the site of inflammation and the main sources include monocytes/macrophages and polymorphonuclear (PMN) leukocytes.

Several studies have found various proteins to be present in different amounts in healthy/gingivitis saliva/gingival tissue (Table 5.1). The majority of literature reports which have attempted to identify biomarkers for gingivitis/periodontitis trials have compared saliva, gingival tissue or gingival crevicular fluid, between groups of pooled 'healthy' versus 'gingivitis' individuals. These studies have often focused on a small subset of enzymes and used antibody detection techniques (e.g. ELISA) [138, 356-358] on account of their ease of use. Many of the procedures used to date estimate the concentration of proteins using a saliva-volume-based normalisation and do not take individual protein concentrations into account. Because the salivary flow rate and protein concentration are known to vary significantly from patient to patient, day to day, as well as the time of day[182, 289], these past studies often suffer from poor resolution and incorrect conclusions may be drawn. There have been

several recent publications within the last year that have used proteomic profiling comparison of saliva from patients with periodontitis and controls using 2D SDS PAGE[359-361]. In these studies, quantitation was based on spot intensity and very few proteins (spots) were found to be upregulated. Only the most abundant proteins were detected using this methodology. For example, one report using 2DGE proteomics, comparing protein abundance changes before and after periodontitis treatment, identified 15 out of 128 protein spots that were differentially changed. Only 10 of these were possible to characterise.[361] These previous reports are often conflicting with regard to the reported changes in protein expression during gingivitis or periodontitis. This is likely a result of the small numbers of patients tested, a pooling of samples, and the utilization of only two time points (before and after periodontitis treatment). Currently, only one large non-gel proteomic comparison of gingivitis/periodontitis has been performed to date[362]. This study analysed proteins in human gingival crevicular fluid in ten people (5 healthy, 5 with gingivitis) and quantitated the protein abundance using label-free technology[362]. A total of 154 proteins were identified, but many of the lower abundance pro-inflammatory cytokines were not detected. In this case, depletion of high abundance proteins could have been performed to identify more protein species. However, recent studies suggest that such depletion methods may unintentionally remove proteins of interest that are complexed to proteins that are being depleted. For example, albumin is known to bind to hormones, cytokines and lipoproteins, and such proteins are inadvertently removed during albumin depletion[363]. Concerns regarding the reproducibility of protein depletion also exist in current literature[362].

Induced gingivitis trials, which potentially allow for the collection of controlled, high quality, longitudinal data, have been very rarely performed in conjunction with biomarker searching. A longitudinal study involves the sampling the same patients over a period of time (during the induction of gingivitis) whilst a cross-sectional study only investigates the subjects at a defined timepoint. The main reported longitudinal trial monitored the transcriptome profile of gingival biopsy tissue from an induced gingivitis trial[364]. However, only 3 samples were taken from each patient (prior to induction, following 28 days of induced gingivitis, and following treatment). Taking samples during the induction of gingivitis would have allowed a time course profile to be generated, which may have substantially improved this study. In addition to gingival tissue sampling, cell culture models of gingival fibroblasts have been investigated and the effect of induction by Interleukin-1β has been used to model the effect of inflammation, effectively modelling gingivitis[365].

The aims of this chapter are to identify putative biomarkers for gingivitis using saliva samples from two clinical trials and to determine whether specific proteomic platforms can provide enough sensitivity to identify previously reported biomarkers that were identified using antibody-based techniques (Table 5.1). In one of these clinical trials, multiple samples were obtained during gingivitis induction to allow for time-course profiling of protein levels. Proteins of both human and bacterial origins are considered for biomarker discovery in the saliva supernatant and residue. Furthermore, to determine whether kinase activity profiling has potential for biomarker discovery in saliva, the application of KAYAK on saliva samples from an induced gingivitis clinical trial are evaluated.

Table 5.1 Proteins reported in the literature that change with the onset of gingivitis/periodontitis. For large studies (e.g. Offenbacher *et al.*[364]), only the top several proteins are reported for each of the literature reports.

Saliva/gingival tissue	Protein (gene symbol)	Protein description	Increased fold	Reference
saliva	PIP	Prolactin	-1.36	Haigh, 2010[361]
	S100A6	Calcium binding	1.64	
	S100A8	Calcium binding	2.31	
	S100A9	Calcium binding	1.45	
	SPLUNC2	Parotid secretory protein	-1.50	
	TALDO1	Transaldolase 1	1.65	
	HP	Haptoglobulin α-chain subunit	3.64	
	GDI2	Rab CDP-dissociation inhibitor	2.17	
	TKT	Transletolase	5.37	
Saliva	GOT	Aminotransferase	Ca. 3.7-8.7	Todorovic, 2006[357]
	LDHA, LDHB	Lactate dehydrogenase	4.7-10.2	
	ALP, ACPP, ANP	Acidic and alkaline phosphatises	1.5-5.3	
	GGT1	Gama glutamyl transferase	1.5-2.6	
Saliva	MMP8	Matrix metalloproteinase 8	4.3	Miller, 2006[358]
	IL1B	interleukin-1 beta	3.5	
Saliva	LDH	Lactate dehydrogenase	1.3-3.2	Zappacosta, 2007[366]
	AST	Aspartate aminotransferase	1.2-2.5	
Saliva	MMP8	Matrix metalloproteinase 8	3.3-4.5	Rai, 2008[356]
Gingival tissue, biospsy	PROK2	Prokineticin 2	4.55	Offenbacher, 2009[364]
	CSF3	Colony stimulating factor 3	2.27	
	CSF2RA	Colony stimulating factor 2 receptor alpha		
	ANGPTL1	Angiopoietin-like	2.07	
	ITGB2	Beta integrin 2	1.84	
	CMTM2	CKLF-like MARVEL TD	1.81	
	IL8	Interleukin 8	1.81	
	IL1A	Interleukin 1 alpha	1.79	
	CCL5	RANTES	1.76	
	IL1B	interleukin-1 beta	1.72	
	CORO1A	Coronin, actin binding protein	1.45	
	SOD2	Superoxide dismutase 2	1.68	
	MUC4	Intestinal mucin	1.94	
Gingival tissues, cell culture	NFKB1	Nuclear factor of kappa light polypeptide gene enhancer in B-cells 1 (p105)	4.86	Vardar-Sengul, 2009[367]
	REL	V-rel reticuloendotheliosis viraloncogene homologue	5.95	
	ATF3	Activating transcription factor 3	4.51	
	CCL2	Chemokine (C-C motif) ligand 2	461	
	CCL5	Chemokine (C-X-C motif) ligand 1	1920	
	MAP3K8	Mitogen-activated protein kinase kinase kinase 8	40.2	
	TNF	Tumour necrosis factor	42.2	

5.2 Experimental procedures

5.2.1 Procter & Gamble gingivitis trial

On wake-up (WU), saliva samples (1 ml) were self-collected by the subject. The saliva was expectorated into a sterile vial and brought by the subject into the P&G site where it was frozen at –80°C. An informed consent form was signed by all participants. In total 20 subjects were selected by the dentist at P&G oral care: 10 displaying some level of gingivitis and 10 showing no symptoms. Subjects were excluded if any of the following criteria applied to them:

- Use of antibiotics, antimicrobial mouthwash, medicated lozenges or chlorhexidine 2 weeks prior to the study or during the study period.
- Sufferers of caries, advanced periodontitis, oral candidiasis or other diseases that could interfere with the study (e.g. diabetes, bacterial infection, malaria, tonsillitis, sinusitis, bronchitis, etc.)
- Have undergone significant oral surgery in the last 12 months or had dental conditions that could interfere with the study.
- Pregnancy or nursing.

Around 2 ml of whole saliva was collected at wake-up from 10 people with gingivitis and 10 healthy people on 6 different days over a 2 week period. The samples were frozen at -80°C prior to processing.

5.2.1.1 Sample preparation and mass spectrometric analysis

Saliva samples from 10 subjects with gingivitis and 10 controls were pooled respectively and the supernatant was analysed by GeLC-MSMS as described in Section 2.3.1 and 2.3.3. In brief, the pooled saliva samples were reduced, alkylated, separated on SDS-PAGE, before being digested in-gel and analysed on an LTQ Orbitrap mass spectrometer at HMS (Section 2.5.1-2.5.3). Proteins were quantitated using spectral counting and the protein abundance was compared between the pooled gingivitis and control set.

In addition to the pooled analysis using GeLC-MSMS, individual saliva samples were analysed individually by in-solution digestion as described in Section 2.2.1 In brief, 20 saliva samples (10 healthy, 10 gingivitis subjects) were reduced, alkylated, digested with trypsin in-solution and analysed by LC-MSMS on the HCT mass spectrometer at Edinburgh (Section 2.4.1.1). Proteins were quantitated using spectral counting and the protein levels compared between subjects with gingivitis and the control subjects.

5.2.2 Eastman Dental Institute induced gingivitis clinical trial

A single centre, randomised, double-blind, and parallel-group (3 groups) clinical trial was performed on 45 subjects (15 per group) who met the all the necessary inclusion and exclusion requirements (Appendix R). Subjects presented to the clinical site for initial screening and enrolment (day -7), having refrained from brushing and inter-dental cleaning for at least 12 hours prior to the baseline visit. Following the screening examinations, the subjects received a complete dental prophylaxis to remove plaque, stains and calculus. Subjects began a regime of twice a day brushing (morning and evening) and daily flossing (evening) for 7 days. On the eighth day, subjects presented to the clinical site for baseline clinical measurements and sample collection (saliva and plaque) and were randomised to their assigned rinse material (day 0). They then began their regimen of rinsing with 10 ml of mouthrinse for 30 seconds twice daily (morning and evening after meals) for 14 days and this was the sole oral hygiene routine (no brushing or flossing). Rinsing was unsupervised and took place at home. Saliva and plaque samples were taken on days 7, 9, 12 and 14. Subjects were randomly assigned to one of three groups with a mouthwash containing 'flavoured water' (negative control), 'chlorhexidine' (positive control) or 'test compound'.

An oral examination was conducted at each visit to monitor the effect of the mouthrinse formulations on the soft and hard tissues. Buccal and sublingual mucosae, lips/labial mucosa, mucobuccal fold, gingiva, tongue, hard and soft palate, uvula, oropharynx, teeth and dental restorations were examined and clinical measurements made. Changes from the baseline and previous visits were recorded at each subsequent clinic visit. Saliva was collected on day 1, 7, 9, 12 and 14 of the trial before any clinical sampling or measurement took place. Study participants refrained from eating, smoking, or drinking for 60 minutes prior to saliva collection. The patient was seated in a quiet environment with their head tilted forward. Whole saliva samples were collected using plastic tubes (approximately 1 ml) to which protease inhibitors were added (Complete protease inhibitor 1x, Roche) prior to freezing at -80°C.

5.2.2.1 Sample preparation

After defrosting the saliva at 4°C, it was centrifuged at 1,500 rpm for 5 min and the residue and supernatant were separated. The supernatant was further clarified by centrifugation at 16,000 rpm and the residue from this second centrifugation step was discarded and the supernatant was used for the KAYAK assay and shotgun proteomic analysis. All centrifugation and preparation was carried out at 4°C to minimise protein degradation. The residue from the first centrifugation step was washed with 1 ml of PBS and lysed using 750 µl of KAYAK lysis buffer (Section 2.7.2). Proteins (20 µg) in the saliva supernatant and residue following lysis were precipitated with methanol/chloroform (Section 2.3.1) after addition of 500 fmol CKM (creatine phosphokinase, rabbit) as an internal standard. The samples were then prepared for shotgun proteomic analysis, as described in Section 2.3.1.

5.2.2.2 Mass spectrometric analysis and data processing

LC-MSMS analysis was performed on an LTQ-Orbitrap Velos mass spectrometer (Thermo Scientific). The LTQ-Orbitrap Velos was operated in data-dependent mode with dynamic exclusion (20 s), where the high resolution survey scan was followed by 20 MS/MS scans in the LTQ on the 20 most abundant ions (TOP 20 method). Apart from the TOP20 acquisition, all other parameters were the same as described in the Experimental Procedures Section 2.5.1 The MSMS data was searched using the SEQUEST algorithm against a combined database of eubacteria, archaea and human (IPI) proteins. Peptides were filtered to a false discovery rate < 1% at the protein level. The protein abundance was estimated by spectral counting normalised to the count of internal standard (CKM, rabbit) peptides. Hierarchical clustering was performed using Clustal 3.0 to group similar responding proteins together.

5.2.3 Kinase activity profiling of saliva

Kinase activity profiling using KAYAK was performed on the saliva supernatant (10 µg) and residue (10 µg) for all samples from the Eastman Dental Institute induced gingivitis trial (60 samples in total). The kinase reactions were performed using the 96 well KAYAK assay format with the K60 peptide substrate set (Section 2.7.5.1).

5.3 Results and discussion

5.3.1 Comparison of healthy and gingivitis saliva samples from Procter & Gamble trial

Wake-up saliva samples were obtained from a P&G trial using 10 healthy and 10 gingivitis sufferers. Over a period of 2 weeks, 6 samples from each patient were obtained. The 6 samples from each patient were pooled to minimise day-to-day variation. Subsequently, the healthy and gingivitis saliva samples from the 10 people in each group were pooled. These 2 supernatant saliva samples were analysed using GeLC-MSMS (10 fractions for each sample) and spectral counting was used for quantitation. In total, 671 proteins (peptide FDR < 0.1%) were identified in the two samples. The spectral counts for the proteins across the 10 fractions for each sample were summed and the fold value (gingivitis/healthy) was determined (Figure 5.2). The results were filtered to consider only those proteins with a total spectral count of greater than 5 and a fold value greater than 2 for gingivitis compared to healthy samples or *vice versa*. The majority of the proteins fall within a fold value of less than two, however, 45 proteins have a spectral count (SC) and fold value which lie beyond this cut-off (Table 5.2). One of the proteins, IL1F6 (interleukin family member 6), is an inflammatory cytokine and would be expected to be unregulated in gingivitis. However, no other cytokines or putative periodontitis biomarkers such as MM8 were significantly upregulated.

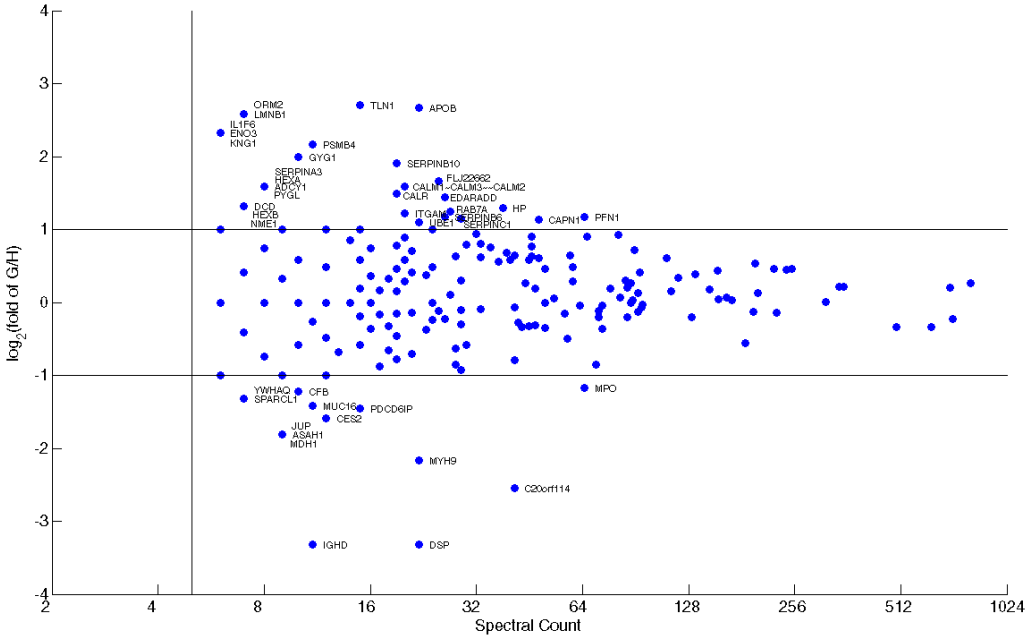


Figure 5.2 Comparison of protein abundance in pooled gingivitis and healthy samples. The fold value of the spectral count for each protein for gingivitis/healthy was plotted against the total spectral count. Proteins with a total spectral count of 5 or less were excluded.

Table 5.2 List of proteins that increase/decrease in gingivitis (G) and healthy (H) pooled saliva samples. The ratio G/H represents the ratio of protein abundance (calculated using spectral counts) of pooled gingivitis *vs.* pooled healthy. Only proteins showing an increase in abundance of greater than 2 fold in either gingivitis (G) or healthy (H) saliva and which had a total spectral count greater than 5 were considered.

Protein	G/H	Spectral count
TLN1	6.5	15
APOB	6.3	22
LMNB1	6.0	7
ORM2	6.0	7
ENO3	5.0	6
IL1F6	5.0	6
KNG1	5.0	6
PSMB4	4.5	11
GYG1	4.0	10
SERPINB10	3.8	19
FLJ22662	3.2	25
ADCY1	3.0	8
CALM1	3.0	20
HEXA	3.0	8
PYGL	3.0	8
SERPINA3	3.0	8
CALR	2.8	19
EDARADD	2.7	26
DCD	2.5	7
HEXB	2.5	7
NME1	2.5	7
HP	2.5	38
RAB7A	2.4	27
ITGAM	2.3	20
PFN1	2.3	65
SERPINB6	2.3	26
SERPINC1	2.2	29
CAPN1	2.2	48
UBE1	2.1	22
MPO	0.4	65
CFB	0.4	10
SPARCL1	0.4	7
YWHAQ	0.4	7
MUC16	0.4	11
PDCD6IP	0.4	15
CES2	0.3	12
ASAH1	0.3	9
JUP	0.3	9
MDH1	0.3	9
MYH9	0.2	22
C20orf114	0.2	41
DSP	0.1	22
IGHD	0.1	11

This result may be expected due to the intrinsic limitations of such a pooled analysis. Whilst the comparison of the pooled healthy and gingivitis samples can give an indication of proteins present in different quantities in saliva from healthy people and those suffering from gingivitis, it takes no account of the variation of protein abundance within each group of healthy and gingivitis people. Taking this into consideration, the individual saliva samples from the healthy and gingivitis people were analysed by LC-MSMS and spectral counting. Due to the large number of samples, it was not practical in terms of instrument time to fractionate each of the samples by SDS-PAGE. Instead, each

sample was subjected to in-solution trypsin digestion prior to analysis by LC-MSMS. A very large amount of variation within the healthy and gingivitis groups was observed and no proteins were significantly different between the healthy and gingivitis group (Figure 5.3). A t-test was performed between the individual samples in the healthy and gingivitis groups but no proteins were found to be consistently significantly changed in abundance (Appendix S)

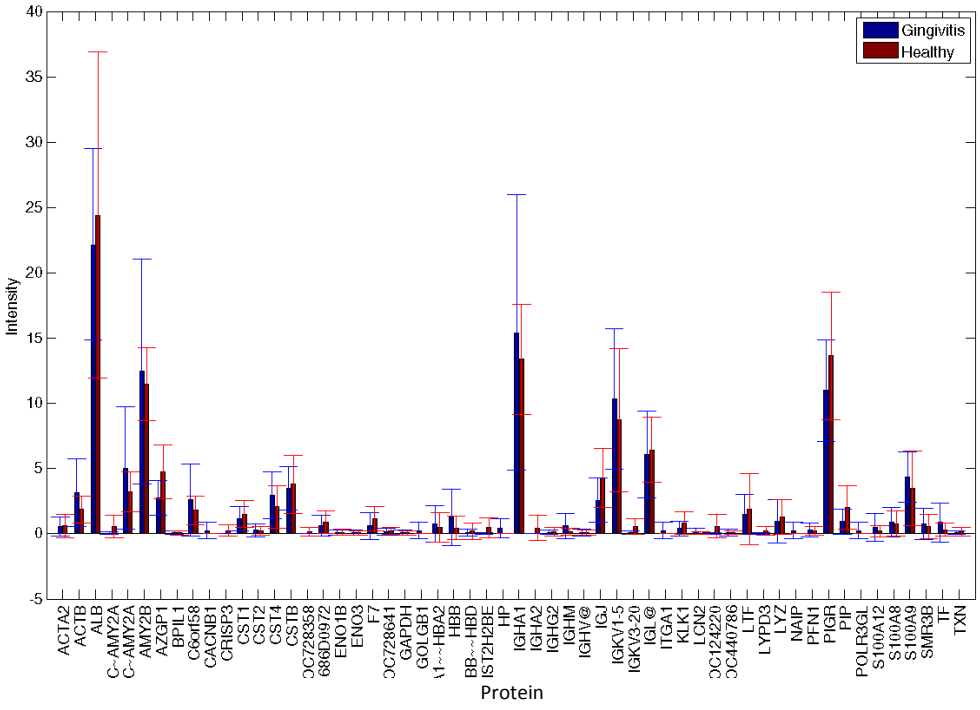


Figure 5.3 Comparison of the mean protein abundance for individual saliva samples from healthy and gingivitis patients. Intensity represents the mean spectral count.

Compared with the experiment where the pooled healthy and gingivitis samples were fractionated by SDS-PAGE, far fewer proteins were identified when the individual healthy and gingivitis samples were analysed by LC-MSMS with no fractionation as a result of increased depth of sampling. Also, the individual healthy and gingivitis samples were analysed on a different LC-mass spectrometer setup (HCT, Bruker) compared to the pooled samples (LTQ, Thermo) and therefore there were two variables.

It was later understood that the saliva samples from the gingivitis and control subjects had not been obtained in the same timeframe, but approximately a year apart. This implies that the trial may have been confounded and that observed differences may be a result of gingivitis or length of storage. As a consequence of this, no in-depth analysis was performed on these samples. However, this cross-sectional study was useful in that it clearly demonstrated that the large inter-person variability must be taken into account to obtain accurate results. Therefore, an effort was directed towards obtaining samples from a longitudinal time course induced-gingivitis clinical trial.

5.3.2 Induced gingivitis clinical trial at the Eastman Dental Institute

Saliva samples were obtained from a longitudinal induced clinical trial performed at the Eastman Dental Institute (London, UK). This was a randomised, controlled clinical trial to determine the effectiveness of a test compound for gingivitis treatment. Saliva samples were collected over a period of time during gingivitis induction (day 1, 7, 9, 12 and 14) where gingivitis induction was through brushing abstention.

For the proof-of-concept work presented here, six samples were obtained (out of the 40 participants). Saliva samples from these six individuals were analysed blinded to the treatment regimen. Patients were either on a daily mouthwash routine of either water (negative control), chlorhexidine (positive control) or a proprietary test compound. Ideally, saliva samples should have been selected after unblinding the randomisation code but this was not possible due to time restrictions. Since the primary aim of this PhD was to detect biomarkers that increase with gingivitis, it was of most interest to analyse the negative control population (patients on the water mouthwash) where the gingivitis symptoms would be expected to increase the most. Due to confidentiality restrictions, it was not possible to include data for patients who were treated with the proprietary test compound.

Both the saliva supernatant and the residue were investigated for each of the 30 samples obtained from six individual subjects on day 1, 7, 9, 12 and 14 of gingivitis induction. The 60 samples (30 supernatant, 30 residue) were analysed by shotgun proteomics and protein abundance was estimated by spectral counting. Sample fractionation (e.g. SDS PAGE) was not practical in terms of available instrument time and each saliva sample was digested in-solution. The whole saliva digest was subsequently analysed in a single 2 hour LC-MSMS run.

For each saliva supernatant sample, around 600 proteins were identified. In total, 977 proteins were identified for all 30 samples combined with a FDR <1 % at the protein level. The total unique peptide count was 6,454 and the total peptide count was 209,539. Of the proteins identified, 894 were human, 79 were from bacteria and 2 were from archaea (Figure 5.4). For each saliva residue sample**,** around 1,000 proteins were identified in each run. In total, 2654 proteins were identified for all 30 samples combined with FDR <1 % at the protein level. The total unique peptide count was 11,807 and the total peptide count was 287,380. Of all the proteins identified, 1378 were human, 1267 were from bacteria and 7 were from archaea. A total of 198 bacteria species were identified (Table 5.3) although many of these may be redundant through the similarity of proteins present in similar species and the incomplete sequencing of bacterial genomes (Appendix T).

Almost half of all the proteins identified in the saliva residue were bacteria proteins. This represents a much greater proportion than that found from the saliva residue from a single healthy individual. presumably on account of the greater population diversity and gingivitis induction.

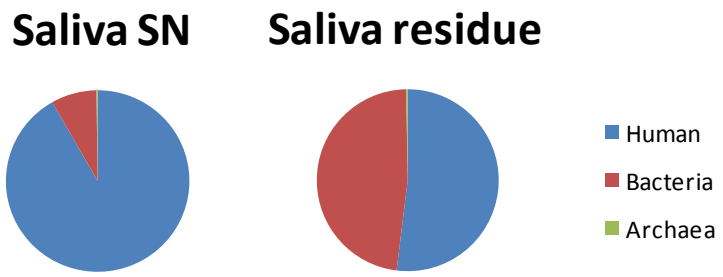


Figure 5.4 Distribution of proteins allocated to human, bacteria and archaea in the saliva supernatant (SN) and residue for all 30 samples combined.

The bacteria from the phyla Proteobacteria (gram-negative) and Firmicutes (mostly gram-positive) dominate in both the saliva residue and the supernatant (Figure 5.5). The genera *Neisseria* and *Streptococcus* are in the phyla Proteobacteria and Firmicutes respectively and have very high spectral counts. Several species in Proteobacteria have been associated with gingivitis/periodontitis.

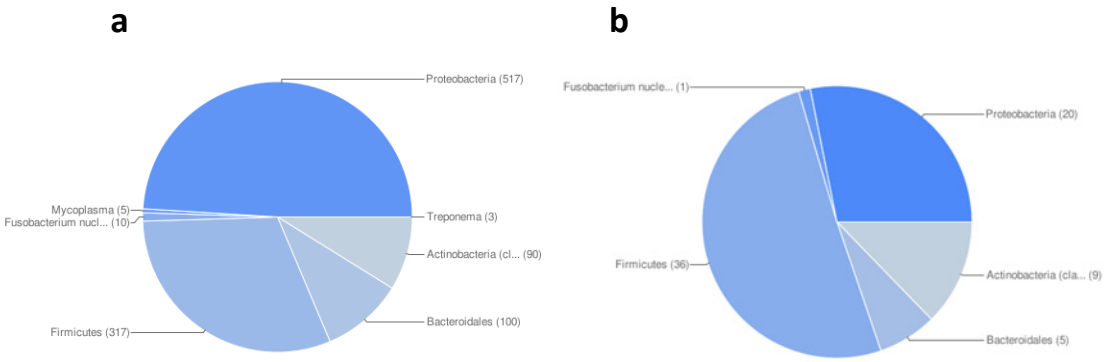


Figure 5.5 Distribution of bacterial proteins for the 30 saliva samples from the induced gingivitis trial across the major phyla identified in saliva (a) residue and (b) supernatant. The pie charts show the number of proteins identified in each bacterial group. The sorting of proteins into the different taxonomic groups was based on the UniProt/Swiss-Prot database annotations.

Table 5.3 A subset of the bacteria from which proteins were identified in the saliva residue and supernatant along with their spectral counts. Only bacterial for which the number of proteins identified for a species was at least 5 are shown. A full list of all the bacteria identified (198 species) is included in Appendix T.

Organism name	Number of proteins		Total Spectral count	
	Residue	Supernatant	Residue	Supernatant
Neisseria meningitidis	121	1	917	1
Neisseria flavescens	95	0	382	0
Neisseria gonorrhoeae	94	2	1026	10
Streptococcus pneumoniae	65	3	780	15
Streptococcus salivarius	34	3	199	69
Haemophilus influenzae	32	0	214	0
Actinomyces odontolyticus	22	2	584	14
Veillonella dispar	21	3	216	29
Bacteroides caccae	17	0	259	0
Streptococcus agalactiae	17	1	296	1
Bacteroides fragilis	16	0	273	0
Selenomonas flueggei	16	2	121	42
Streptococcus gordonii	15	0	291	0
Haemophilus ducreyi	14	0	46	0
Streptococcus thermophilus	14	2	119	3
Streptococcus sanguinis	13	4	224	47
Bacteroides vulgatus	12	0	187	0
Lactococcus lactis	12	1	158	1
Streptococcus mutans	12	0	135	0
Streptococcus equi	11	1	62	2
Fusobacterium nucleatum	10	1	91	9
Porphyromonas gingivalis	10	1	115	5
Streptococcus pyogenes serotype	10	1	111	1
Acinetobacter baumannii	9	1	111	18
Escherichia coli O139:H28	9	1	221	6
Kocuria rhizophila	9	0	95	0
Actinomyces urogenitalis	8	1	70	1
Bifidobacterium longum	7	1	12	3
Eikenella corrodens	7	0	43	0
Eubacterium ventriosum	7	1	133	4
Kingella oralis	7	0	16	0
Ralstonia solanacearum	7	0	24	0
Streptococcus infantarius	7	0	39	0
Burkholderia cenocepacia	6	0	37	0
Enterococcus faecalis	6	0	37	0
Porphyromonas uenonis	6	0	26	0
Bacteroides coprocola	5	0	42	0
Bacteroides thetaiotaomicron	5	2	62	38
Bifidobacterium animalis	5	1	98	21
Bordetella avium	5	1	139	15
Delftia acidovorans	5	1	58	24
Ralstonia pickettii	5	0	24	0
Streptococcus equinus	5	1	75	13
Streptococcus suis	5	2	55	12

5.3.2.1 Induced gingivitis: spectral count comparison of human proteins

Previous literature reports comparing saliva from healthy individuals to that from gingivitis/periodontitis patients have reported particular proteins that were found to be higher in gingivitis/periodontitis compared to healthy. To determine whether these proteins were present in increasing quantities during gingivitis induction, the spectral count profile of particular proteins was plotted (Figure 5.6) and a clear upward trend was observed (t-test, $p<0.05$), corroborating these past results.

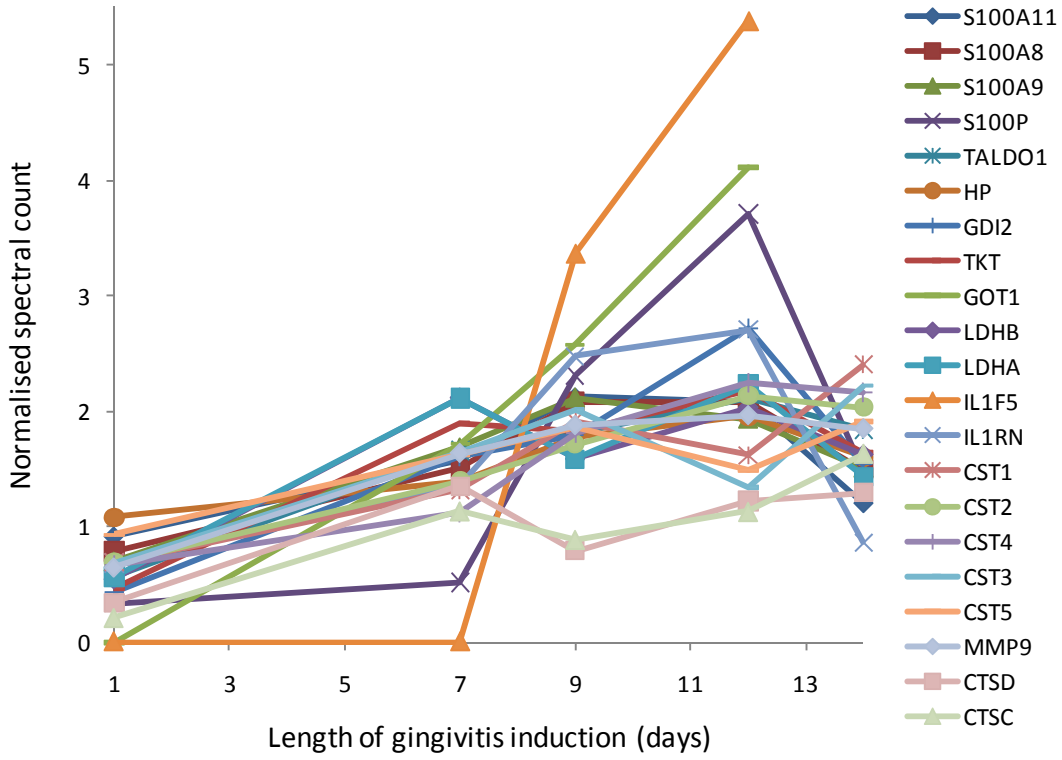


Figure 5.6 Profile of protein abundance (spectral count) of various proteins with period of gingivitis induction in the saliva supernatant from patient 1 (water mouthwash). Spectral count is normalised to the spectral count of the internal standard protein (Rabbit CKM) followed by normalisation of the sum of the spectral counts across the all the days of gingivitis induction and gives an estimation of protein abundance. Selected proteins were those reported to be present in greater quantities in patients with gingivitis/periodontitis.

Several notable proteins which have been previously associated with gingivitis show upward trends upon gingivitis induction. First, an increase in the abundance of cystatin proteins was observed in this study. For cystatin 4, a large increase was observed in the later part of gingivitis induction (day 7 to 14), whilst for the other cystatins (CST1-3, 5) a more linear trend was observed except for a slight

decrease on day 12. Cystatins are cysteine proteinase inhibitors that reduce tissue damage in inflammation and are elevated in patients with gingivitis/periodontitis[368-369].

Also, the abundance of MMP9 (matrix metalloproteinase 9) was found to increase with gingivitis induction, and whilst MMP8 was identified, the spectral count was low and accurate quantitation was not possible. This is consistent with published data linking MMP8 and MMP9 to connective tissue breakdown in gingivitis[356]. MMPs are zinc dependent endopeptidases associated with polymorphonuclear neutrophils that are known to be expressed in many inflammatory diseases[362] The abundance of cathepsin (CTS) proteins C, D and G was increased in the saliva supernatant, which is consistent with previous finding[370-372]. Cathepsins, along with MMPs are proteases associated with polymorphonuclear neutrophils (PMNs)[371].

Various cytokines were also identified including interleukins. The cytokine IL1F5 increased dramatically during the second half of gingivitis induction (day 7 to 14). Whilst various annexins were identified, the spectral count was too low for accurate quantitation and no significant trend was observed. Annexin is another member of the calcium binding family that is involved with the defense and inflammatory response[373]. Annexin1 is thought to be a negative regulator of various pro-inflammatory mediators including cyclooxygenase, IL-1 and IL-6[374]. The balance between protease activity and inhibition is essential for the maintenance of healthy tissues. Increased protease activity and/or decreased inhibition may result in tissue inflammation and ultimately tissue destruction.

There was a clear, consistent increase in S100 protein abundance during gingivitis induction. S100 proteins are calcium binding proteins that are expressed in inflamed epithelial cells and macrophages and decrease in saliva following treatment of periodontitis[375]. Therefore, the increased abundance of S100 proteins with gingivitis induction in this study is consistent with the expected response to gingivitis.

Haptoglobulin (HP), an acute phase protein, was increased in the gingivitis samples, which is consistent with previous work where increases in HP were observed in saliva and in the serum of periodontitis patients[376]. Transketolase (TKT) and transaldolase (TALDO1) were also increased with the duration of gingivitis induction. Whilst TKT and TALDO1 are not known to have a direct role in the immune response, a previous proteomics study comparing the salivary proteome of individuals before and after periodontal treatment found these proteins to be higher in the diseased state[361]. Glutamic-oxaloacetic transaminase (GOT), gama glutamil transferase (GGT) and lactate dehydrogenase (LDH) are enzymes associated with cell injury and cell death and were found to increase with gingivitis induction, consistent with previous reports[357].

The majority of literature studies on the investigation of gingivitis/periodontitis biomarkers have been performed on saliva, with a few have been performed on gingival tissue. To the author's knowledge, no studies have been published on analysis of the saliva residue with respect to gingivits. However, because a significant portion of saliva residue contains sloughed-off and lysed gingival cells, the

analysis of saliva residue would likely give similar results to direct analysis of gingival tissue, which is more difficult to obtain. The profile of proteins in the saliva residue which have previously been reported to be higher in gingival tissue and saliva are considered in Figure 5.7.

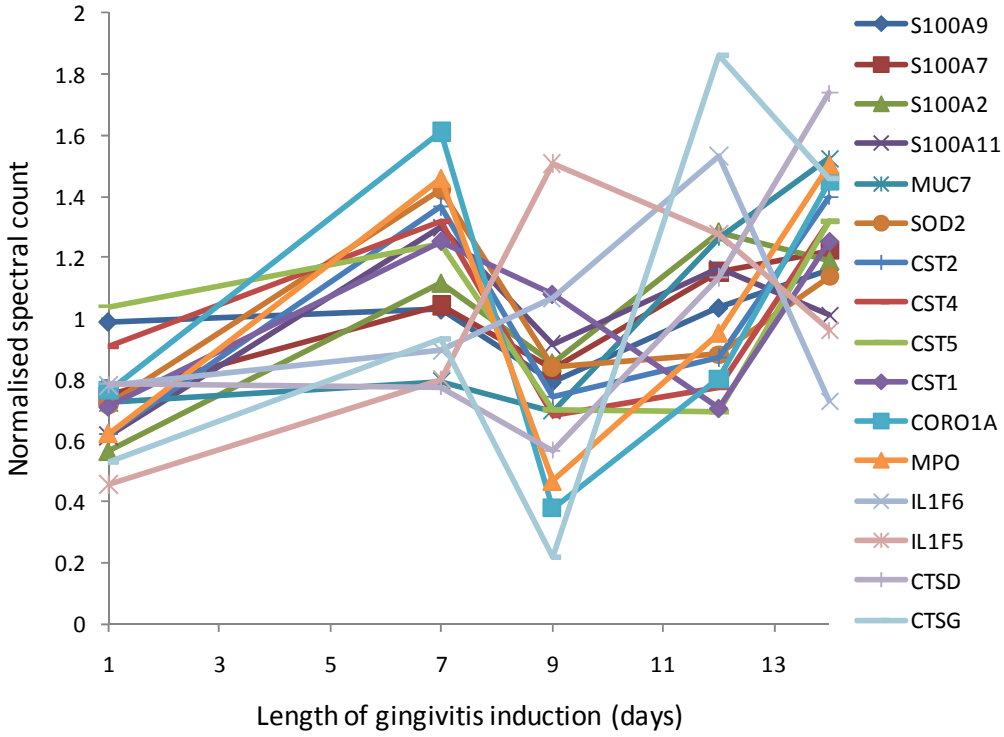


Figure 5.7 Profile of protein abundance (spectral count) of selected proteins with length of gingivitis induction for subject 1 (water mouthwash) in the saliva supernatant. Spectral count is normalised to the spectral count of the internal standard protein (Rabbit CKM) followed by normalisation of the sum of the spectral counts across the all the days of gingivitis induction and gives an estimation of protein abundance. Selected proteins taken from literature where they were reported to be higher in patients with gingivitis/periodontitis.

The correlation of proteins reported to be increased in gingivitis/periodontitis with those found in the saliva residue appears poorer than the correlation for the proteins in the saliva supernatant (Figure 5.6 and Figure 5.7). This may be due to several factors. Firstly, comparing literature 'saliva supernatant biomarkers' with saliva residue is not a fair comparison and many of the proposed biomarkers identified in gingival tissue using transcriptomics (such as PROK2, CSF3, IL8) were either not identified in the proteomic analysis of the saliva residue or were too low abundance for accurate quantitation. Secondly, a significant proportion of the saliva residue comprises bacterial cells, and therefore the abundance of a particular protein is measured against the background of total human and total bacterial proteins (Figure 5.8b, discussed in Section 5.3.2.2). While a particular human protein may increase with induction of gingivitis, it may appear to decrease with respect to the overall number of proteins present due to large changes in bacterial protein concentrations. However, despite variability throughout the course of gingivitis induction, a general upward trend is still observed for

certain proteins reported to be associated with gingivitis including cystatins (CST1, 2, 4 and 5), cathepsins (CTSD and CTSG) and interleukins (IL1F5, IL1F6).

Whilst many of the previously reported biomarkers for gingivitis/periodontitis increase with the induction of gingivitis in these data, several inconsistencies are present. For example, neutrophil defensin 1 (DEFA1) was observed to increase in both the saliva supernatant and the saliva residue (data not shown). Defensins are cytotoxic peptides and are abundant in neutrophil granules in the epithelia of mucosal surfaces, and it has previously been postulated that the progression of gingivitis/periodontitis may be linked to a decrease in the defensin barrier[377-378]. However, DEFB4 was previously reported to be downregulated in gingival tissue in transcriptome pattern profiling of gingivitis induction[364]. The inconsistency with literature could be due to the different sample type (saliva residue *vs.* gingival tissue) along with the low statistical power of this study.

Despite this, during the course of gingivitis induction, the majority of the proposed biomarkers are shown to increase. These results provide evidence that our techniques are valid and may be used as a tool for further gingivitis biomarker discovery. Despite the lack of statistical power in our data, this proof-of-concept work illustrates several potentially novel protein indicators of gingivitis.

Many new proteins not previously reported show correlations with the progression of gingivitis (Table 5.4). Although more patient data is clearly necessary for any reliable conclusions to be drawn, these data suggest that the combination of a longitudinal clinical trial with high throughput proteomic analysis is a powerful tool for disease biomarker discovery. Most significantly, many more proteins were identified and quantified in this study than in previous studies and proteomic comparisons[358, 360-362, 366]. This technique would increase the number of candidate proteins associated with gingivitis and allow for high-throughput screening of potential biomarkers.

Several notable proteins increase with gingivitis induction and have not previously been reported. In the saliva supernatant, Cysteine-rich secretory protein 3 (CRISP3) was observed to have a high correlation coefficient with the length of gingivitis induction. CRISP3 is present in neutrophilic granulocytes and exocrine secretions and is thought to play a role in the innate host defense[379]. CRISP3 has previously not been associated with gingivitis/periodontitis but was shown to be downregulated in patients with Sjogren's syndrome[380] and in patients with oral tongue squamous cell carcinoma[381]. CRISP3 has also been associated with prostate cancer[382] and chronic pancreatitis, a progressive inflammatory process, where increased levels were observed in chronic pancreatitis tissue compared to normal pancreatic tissue[383]. The exact function of CRISP3 is unknown but it has been postulated to act as an anti-infection agent[383] and therefore it might be expected to increase with during gingivitis.

Calumenin (CALU) is a calcium-binding protein located in the endoploasmic reticulum secretory pathway. Calumenin has been implicated in several diseases including cardiovascular, cancer and neuromuscular[384] but the *in vivo* function has been little studied. Data from a study of calumenin in

Caenorhabditis elegans suggests that it is important in calcium signalling pathways[385] and is upregulated in 'stress' conditions[384] which may be present in gingivitis.

Endoplasmic reticulum oxidoreductin 1-L (ERO1L) is thought to be involved in the formation of hydrogen peroxide in the endoplasmic reticulum[386] and is associated with endoplasmic reticulum stress-induced apoptosis where knock down of ERO1L in endoplasmic reticulum in stressed cells was shown to suppress apoptosis [387]. Although ERO1L has previously not been associated with gingivitis/periodontitis, endoplasmic reticulum stress-induced apoptosis is involved in many diseases including diabetes, atherosclerosis and neurodegenerative disease[387]. Inflammation in gingitivitis/periodontitis is known to be associated with activation of the NF-κB pathway and antiapoptotic genes[367]. Therefore, an association of ERO1L and gingivitis/periodontitis may be possible.

Glyceraldehyde-3 phosphpate dehydrogenase (GAPDH) was shown to be activated in T lymphocytes which are involved in adaptive immunity by interacting with the antigen and secreting cytokines[388]. Whilst no previous reports have highlighted the link between GAPDH and gingivitis/periodontitis, an activation of T cells is known to occur in gingivitis inflammation. Glia maturation factor-gamma (GMFG) has been shown to be preferentially expressed in inflammatory cells[389] and its increased abundance with gingivitis induction is logical.

A variety of immunoglobulins were observed to increase with gingivitis induction, where IGHA1, IGHV1OR15-1 and IGJ had high correlation coefficients. In addition, IGLV7-46, IGKV3-20, IGLL1, IGLV2-18, IGHG2, IGKV3D-11, IGKV3D-15, IGKV4-1, IGL, IGLV3-25 all showed positive correlation with gingivitis induction. Previous studies have reported a higher abundance of immunoglobulins in periodontitis saliva and/or gingival crevicular fluid[194, 359-360, 390-391]. This is consistent with the intuitive notion that inflammation of gingival tissues leads to migration of leukocytes and may lead to increased immunoglobulin secretion.

In the saliva residue, leukotriene A4 hydrolase (LTA4H) and Peptidase inhibitor 3 (PI3) both had good correlation with the length of gingivitis induction (correlation coefficient of 1, $p<0.02$). LTA4H is an enzyme that is involved in inflammation and catalyses the synthesis of leukotriene B4, a potent proinflammatory lipid mediator[392]. PI3 is an antimicrobial peptide against both gram-positive and gram-negative bacteria and is known to be upregulated by bacterial lipopolysaccharide and cytokines. Despite no previous association of either LTA4H or PI3 with gingivitis/periodontitis, their increased abundance with increased inflammation of the gingiva would appear logical and these could be promising indicator of gingivitis severity in the saliva residue.

The use of single mass spectrometry runs on in-solution digests of single samples has enabled the profiling of many proteins during the course of gingivitis induction. Spectral counting was used as a means of quantitation. To enable more accurate quantitation of lower abundance proteins, an isotopic labelling method (e.g. dimethylation) would be appropriate. Despite poor statistical power and the

clear need to investigate a greater number of subjects, several biomarkers previously proposed in the literature were found to increase with gingivitis induction, providing support for the presented method. Several proteins, previously not associated with gingivitis/periodontitis, show good correlation with length of gingivitis induction and merit further investigation. These proteins include CRISP3, CALU and ERO1L in the saliva supernatant and LTA4H and PI3 in the saliva residue.

Table 5.4 List of proteins that show positive correlation with the length of gingivitis induction in the supernatant and residue. Correlation using Spearman rank was used and the associated correlation coefficient and p value are shown. The top 100 proteins for the saliva supernatant and residue are shown.

Protein in supernatant	Correlation coefficient	p value
CRISP3	1.00	0.02
CALU	0.97	0.03
ERO1L	0.97	0.03
GAPDHS	0.97	0.03
GMFG	0.97	0.03
PODXL	0.97	0.03
HMG1L10	0.90	0.08
IGHA1	0.90	0.08
IGHV1OR15-1	0.90	0.08
IGJ	0.90	0.08
LOC390006	0.90	0.08
PEBP1	0.90	0.08
PSMA2	0.90	0.08
RETN	0.90	0.08
S100A12	0.90	0.08
S100A4	0.90	0.08
S100P	0.90	0.08
ECGF1	0.87	0.07
TXN	0.87	0.07
ACTA1	0.80	0.13
ACTR2	0.80	0.13
PSMA3	0.80	0.13
SLPI	0.78	0.20
SPRR2C	0.78	0.20
YWHAE	0.78	0.20
ANXA5	0.71	0.40
CD55	0.71	0.40
IGLV7-46	0.71	0.40
SDCBP	0.71	0.40
ADSS	0.70	0.23
CD9	0.70	0.23
CFL1	0.70	0.23
CST2	0.70	0.23
CST3	0.70	0.23
CST4	0.70	0.23
CTSC	0.70	0.23
DKFZ	0.70	0.23
DMBT1	0.70	0.23
EDARADD	0.70	0.23
FAM3D	0.70	0.23
FOLR1	0.70	0.23
HMGN2	0.70	0.23
IGKV3-20	0.70	0.23
IGLL1	0.70	0.23
IGLV2-18	0.70	0.23
KLK1	0.70	0.23
LOC729708	0.70	0.23
MMP9	0.70	0.23
MSN	0.70	0.23

Protein in residue	Correlation coefficient	p value
ARPC2	1.00	0.02
BRPF1	1.00	0.02
LTA4H	1.00	0.02
PI3	1.00	0.02
PSMA4	1.00	0.02
YWHAE	1.00	0.02
TCN1	0.97	0.03
AHCY	0.90	0.08
ALDH3A1	0.90	0.08
ATP6V1B2	0.90	0.08
CES2	0.90	0.08
DDAH2	0.90	0.08
DSC2	0.90	0.08
GAPDH	0.90	0.08
GDI2	0.90	0.08
GPI	0.90	0.08
GSS	0.90	0.08
HSPA4	0.90	0.08
LCN1	0.90	0.08
LOC643997	0.90	0.08
LRRFIP1	0.90	0.08
MAPK1	0.90	0.08
MDH2	0.90	0.08
NP	0.90	0.08
PDIA6	0.90	0.08
PFN1	0.90	0.08
PGD	0.90	0.08
PGLS	0.90	0.08
PSMA1	0.90	0.08
PSMA3	0.90	0.08
REEP5	0.90	0.08
RPLP1	0.90	0.08
S100A7	0.90	0.08
SPINK5	0.90	0.08
FLJ25715	0.87	0.10
LAD1	0.87	0.10
PFDN2	0.87	0.10
SCAMP2	0.87	0.10
ACPP	0.80	0.13
CAP1	0.80	0.13
CAT	0.80	0.13
DECR1	0.80	0.13
DSTN	0.80	0.13
ENO1	0.80	0.13
ENO2	0.80	0.13
ETFA	0.80	0.13
GLOD4	0.80	0.13
IGKV4-1	0.80	0.13
KLK7	0.80	0.13

MYL6	0.70	0.23
NPC2	0.70	0.23
PAM	0.70	0.23
PFN1	0.70	0.23
PGAM4	0.70	0.23
PIGR	0.67	0.27
PSMA5	0.67	0.27
RAB10	0.67	0.27
RDX	0.67	0.27
SERPINB1	0.62	0.30
SERPINB3	0.62	0.30
ARPC1B	0.60	0.35
AZGP1	0.60	0.35
C6orf58	0.60	0.35
CD59	0.60	0.35
CORO1A	0.60	0.35
CST1	0.60	0.35
CYBB	0.60	0.35
FABP5	0.60	0.35
GDI2	0.60	0.35
GRN	0.60	0.35
HDGF	0.60	0.35
HP	0.60	0.35
IGHG2	0.60	0.35
IGKV3D-11	0.60	0.35
IGKV3D-15	0.60	0.35
IGKV4-1	0.60	0.35
IGL	0.60	0.35
IGLV3-25	0.60	0.35
KLK11	0.60	0.35
KRT75	0.60	0.35
LGALS3BP	0.60	0.35
LOC440786	0.60	0.35
LOC643997	0.60	0.35
M6PRBP1	0.60	0.35
MFI2	0.60	0.35
RNASET2	0.60	0.35
SPARCL1	0.60	0.35
TALDO1	0.60	0.35
TPM4	0.60	0.35
VIL2	0.60	0.35
WFDC2	0.60	0.35
LTA4H	0.56	0.40
NCF4	0.56	0.40
NME1	0.56	0.40
P4HB	0.56	0.40
ALDOA	0.50	0.45
AMY1A	0.50	0.45
BPIL1	0.50	0.45
C7orf24	0.50	0.45
CALR	0.50	0.45

MUC2	0.80	0.13
NDUFB10	0.80	0.13
NIPSNAP3A	0.80	0.13
NPEPPS	0.80	0.13
PGAM4	0.80	0.13
PGK1	0.80	0.13
PITPNB	0.80	0.13
PLS3	0.80	0.13
PPP2R1A	0.80	0.13
PRDX1	0.80	0.13
S100A2	0.80	0.13
SOD1	0.80	0.13
SPRR1B	0.80	0.13
ST13	0.80	0.13
TGM3	0.80	0.13
TPM4	0.80	0.13
TUBB2C	0.80	0.13
VIL2	0.80	0.13
CTSA	0.71	0.40
NDUFA11	0.71	0.40
PKP1	0.71	0.40
PLTP	0.71	0.40
PLUNC	0.71	0.40
PRH2	0.71	0.40
THBS1	0.71	0.40
ACTN4	0.70	0.23
ACTR2	0.70	0.23
ALDOA	0.70	0.23
ANXA4	0.70	0.23
ARHGDIB	0.70	0.23
ARPC3	0.70	0.23
ARPC5	0.70	0.23
AZU1	0.70	0.23
CALML3	0.70	0.23
CAPG	0.70	0.23
CFB	0.70	0.23
CS	0.70	0.23
CST2	0.70	0.23
CYC1	0.70	0.23
DDAH2	0.70	0.23
DMBT1	0.70	0.23
DSG1	0.70	0.23
ELA2	0.70	0.23
EPX	0.70	0.23
F3	0.70	0.23
HSPA8	0.70	0.23
HSPD1	0.70	0.23
IDH1	0.70	0.23
IGHA1	0.70	0.23
IGJ	0.70	0.23
IVL	0.70	0.23

5.3.2.2 Induced gingivitis: spectral count comparison of bacterial proteins

Various bacterial species are associated with gingivitis and periodontitis. As well as considering the human portion of the saliva samples from the induced gingivitis trial, the change in abundance of bacterial proteins and species was also investigated. The majority of research has concentrated on identifying periodontal pathogens using classical culturing techniques or using targeted DNA approaches on small sets of putative periodontal pathogens[393, 394-396]. Most studies analysing mouth bacteria have used tooth biofilm (plaque). This is the first study to investigate the proteomic composition of bacteria in saliva in health and disease states.

Samples were analysed by correlating the abundance of each bacterial protein to the length of gingivitis induction. Many changes in bacterial protein abundance over the course of the induced gingivitis trial are observed (Figure 5.8). Notably, the percentage of bacterial proteins in the saliva residue increases dramatically on the last day of gingivitis induction (day 14).

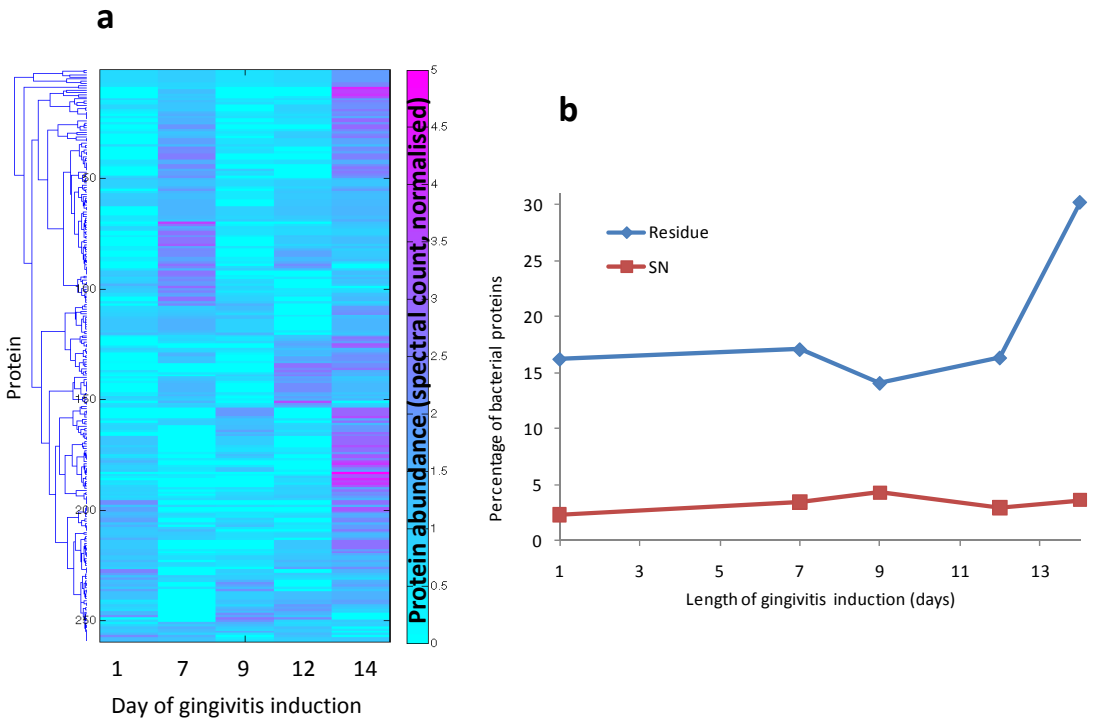


Figure 5.8 Distribution of bacterial proteins with gingivitis induction. Heat map of bacterial protein abundance in saliva residue for the progression of gingivitis for patient 1 (water mouthwash). Each protein is normalised to the internal standard protein (rabbit KCM), added prior to protein precipitation, followed by normalisation to the sum of the spectral counts across all days of gingivitis induction. Only proteins that had a total spectral count of at least 10 were included. (b) Percentage of proteins observed in the saliva residue and supernatant from bacteria. A large increase in the percentage of bacterial proteins was observed on the last day of gingivitis induction.

Various bacterial species and genera have been reported to be associated with periodontitis including *Eubacterium*[397], *Eikenella corrodens*[393], *Bacteroides*[397], *Treponema*[398], *Fusobacterium nucleatum*, *Streptococcus mutans*[393] *and Porphyromonas gingivalis*[399]. As an estimate of the abundance of the species and genera, the number of proteins identified for each general/species was used to profile the abundance of the bacterial genera/species with the progression of gingivitis. No normalisation of the data was performed since 'protein count' rather than spectral counting was used as a measure of abundance. Whilst there is variation with the progression of gingivitis, a slight increase in a few putative periodontal pathogens was observed (Figure 5.9). In previous literature, biofilms of dental plaque are used and therefore larger changes might be expected when compared to saliva. Sampling of the bacterial biofilms can allow a more complete picture of the microbial ecology whilst analysis of the bacteria from saliva is probably only sampling a portion of the bacteria and may underestimate the abundance of bacteria present deep in the biofilm/plaque or those which adhere strongly to mouth surfaces. Despite this, a positive correlation of many of the periodontal pathogens was obtained.

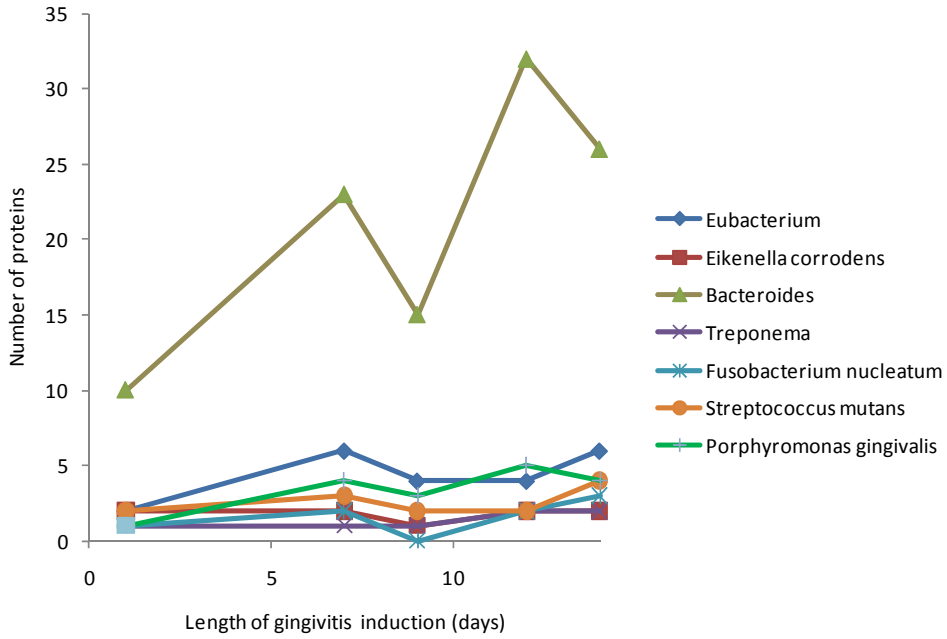


Figure 5.9 Profile of bacterial abundance with progression of gingivitis. The number of proteins matched to bacterial species or genera is used as a measure of protein amount. No normalisation of the data was performed since 'protein count' rather than spectral counting was used as a measure of abundance. Bacteria species/genera were selected from proposed putative periodontal pathogens. Data are from patient 1 (water mouthwash)

The abundance of certain individual bacterial species often rose in parallel with the length of the gingivitis induction period (Figure 5.10). Many bacterial species were observed to increase with gingivitis progression and, in many cases, species from the same genera cluster together: in particular the genera *Neisseria*, and *Streptococcus*. Proteins showing high correlation from the cluster plot are

plotted in Figure 5.11. Of particular interest, the species *Haemophilus influenzae* and *Haemophilus. ducreyi* show good correlation with induced gingivitis. The *Haemophilus* genera is in the Gram-negative proteobacteria phyla and is known to be a key bacterium in the development of bacterial biofilms[400-401]. *H. influenzae* is known to be found in the upper respiratory system[402] and can lead to pneumonia and bacterial meningitis, amongst a wide range of diesases. Whilst *Haemophilus* species have not been reported as a periodontal pathogen, *H. influenzae* has been found in the periodontal pockets in immunocompromised patients[403], and elderly patients with periodontitis are more susceptible to aspiration pneumonia from *H. influenza* and *S, pneumonia* infections[402]. The level of *H. influenza* may therefore be expected to increase with the induction of gingivitis. Interestingly, *S. pneumoniae* was also observed to increase with the induction of gingivitis.

Various species of *Bacteroides* increased during gingivitis induction including *B. caccae*, *B. vulgatus*, *B. coprocola*. *Bacteroides* species are anaerobic and have previously been associated with periodontitis, particularly *B. forsythus*[216]. *Porphyromonas gingivalis* was previously classified in the *Bacteroides* genera[404] and is known to be a key bacterium in the development of periodontitis[405-407]. Very few studies have been performed on bacterial profiling of patients with gingivitis, but *P. gingivalis* is reported to have good correlation with the severity of periodontitis[393] and therefore a small increase with gingivitis induction may be expected. There are no previous reports linking *B. caccae*, *B. vulgatus* or *B. coprocola* with gingivitis/periodontitis and the species are generally associated with the intestine and faecal samples[408-409].

Fusobacterium nucleatum was observed to increase with gingivitis induction and is known to be involved with dental plaque formation on account of their coaggregation with other oral bacteria. Moreover, *F. nucleatum* has previously been reported to be a key bacterium in periodontitis [216, 287]. *Fusobacterium* are gram-negative anaerobes and are involved in the activation of neutrophils[410].

Whilst the mouth is not a natural habitat for *E. coli* and there are no links with gingivitis, several reports have identified the presence of *E. coli* in periodontal pockets[396, 411-413]. *N. flavescens* increases to a great extent on day 14 of gingivitis induction. *Neisseria* are commensal, Gram-negative bacteria that colonise mucosal surfaces[410, 414] and have not been previously associated with gingivitis/periodontitis.

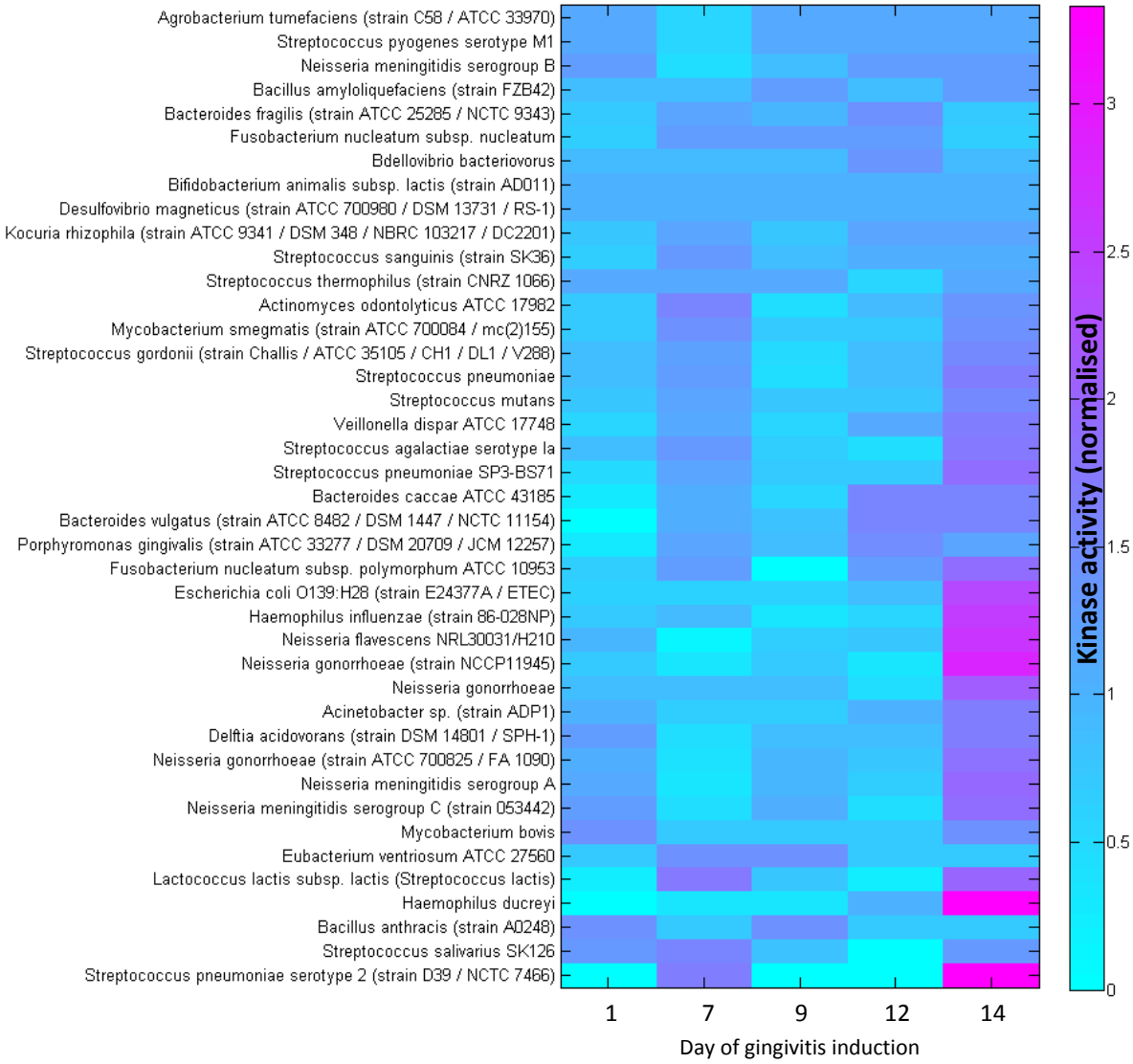


Figure 5.10 Heat map of bacteria in the saliva residue for patient 1. The number of proteins identified for a particular species is used to estimate the abundance of the species. Each row (bacterial species) is normalised to the sum of the proteins across all days of gingivitis induction. Only bacterial species in which 7 proteins or more are identified across all the days of gingivitis for patient are included. Data are for patient 1 (water mouthwash).

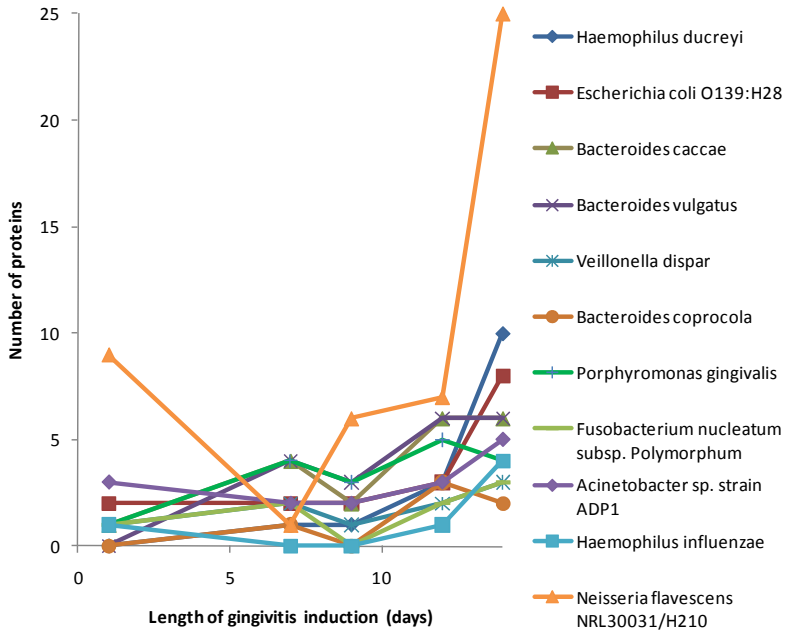


Figure 5.11 Profile of bacterial abundance with progression of gingivitis. The number of proteins matched to bacterial species or genera is used as a measure of protein amount. No normalisation of the data was performed since 'protein count' rather than spectral counting was used as a measure of abundance. Bacteria species/genera were selected from the cluster diagram in Figure 5.10 which had good correlation with the length of gingivitis induction. Data are from patient 1 (water mouthwash).

One potential drawback of this method is that the accurate profiling of bacterial abundance in saliva using the single mass spectrometry run in combination with protein number as a measure of bacteria abundance suffers from low numbers of proteins identified per species. This may be remedied in the future by using spectral counting or isotopic labelling.

Despite this potential drawback, this proof-of-concept work to investigate bacterial proteins in saliva with induced gingivitis has shown that it is possible to profile the abundance of many bacteria in a single 2-hour mass spectrometry run per sample. Various bacteria known to be present in greater abundance in gingivitis/periodontitis were also increased with the length of gingivitis induction in these data. This finding validates this method of quantitating bacterial protein populations in saliva and suggests such a method could have future potential in bacterial biomarker discovery. The numerous bacteria species correlated with gingivitis induction in these data, but not previously reported, suggest that the proliferation and interaction of bacteria with gingival tissues may be more complex and widespread than previously believed. The accurate profiling of bacterial abundance in saliva using the single 2 hour mass spectrometry run suffers from low numbers of proteins identified per organism. Alternative ways of analysis could include spectral counting and isotopic labelling, or sample fractionation.

5.3.2.3 Induced gingivitis: kinase activity profiling of saliva samples

Traditional proteomics focuses on protein abundance quantitation. Protein activity levels, as discussed in Chapter 4, complement the quantitation of protein abundance and provide an additional avenue of exploration. Kinase activity was previously shown to be present in saliva (Chapter 4) and the saliva samples from the induced gingivitis trial reported here were subsequently analysed using KAYAK to obtain kinase activity profiles for both the saliva supernatant and residue (Figure 5.12).

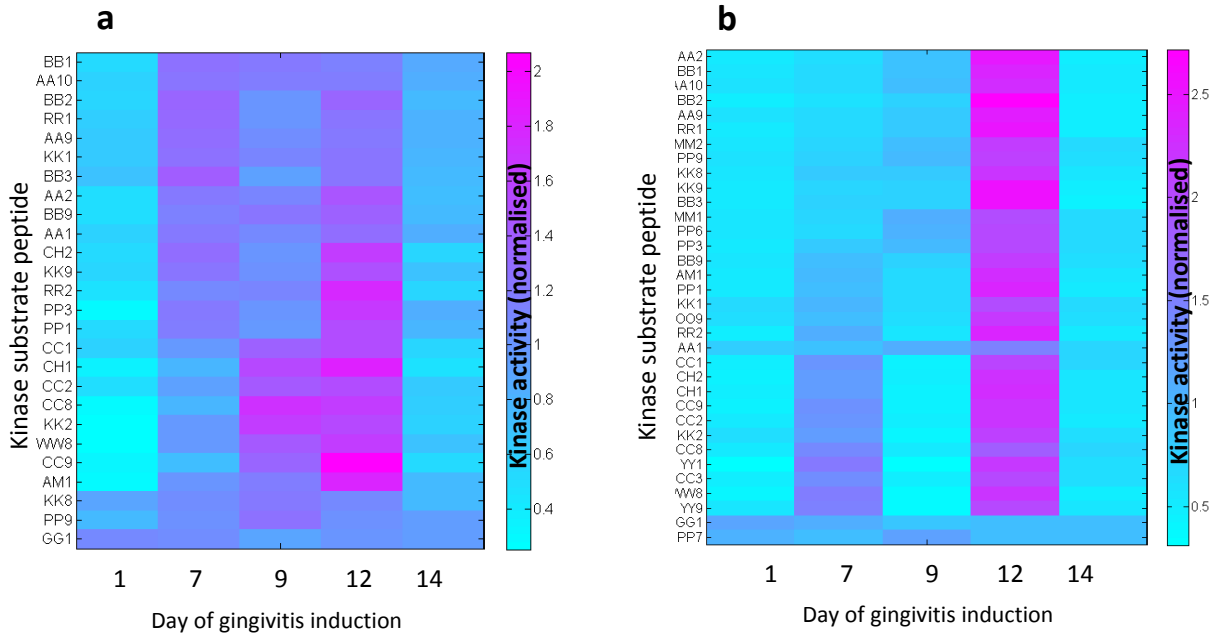


Figure 5.12 Heatmap of KAYAK activities in saliva (a) supernatant and (b) residue for day 1, 7, 9, 12 and 14 of gingivitis induction. Analysis by Pearson correlation hierarchical clustering was used to group similar responders together. The activity of each peptide was normalised to the sum of the kinase activities over all the days. A general increase in kinase activity was observed from day 1 to 12 of gingivitis induction with a large decrease on day 14. Data are from patient 1 (water mouthwash).

A general trend of increased kinase activity with the induction of gingivitis was observed in both the saliva supernatant and residue. A large decrease in activity was observed on the last day of gingivitis induction (day 14). A possible reason for this decrease is that because of the rapidly accelerating gingivitis severity in this final sampled day, a greater proportion of the sample was comprised of bacterial proteins. Previous proteomic analysis (Section 5.3.2.2) showed that the number of bacterial proteins identified on day 14 in the saliva residue was twofold higher than previous samples (Figure 5.8). Therefore, the lysate on day 14 would be expected to contain a lower proportion of human proteins. Assuming the kinase assay only measures the human kinase activities with no contribution from bacterial kinases, this observed decrease is logical. The decrease in kinase activity on day 14 was particularly evident in the saliva residue, consistent with the large increase in proportion of proteins matched to bacteria. To allow easier comparison of the different kinase profiles, day 14 was omitted in the final analysis and days 1 to 12 were profiled and clustered together (Figure 5.13).

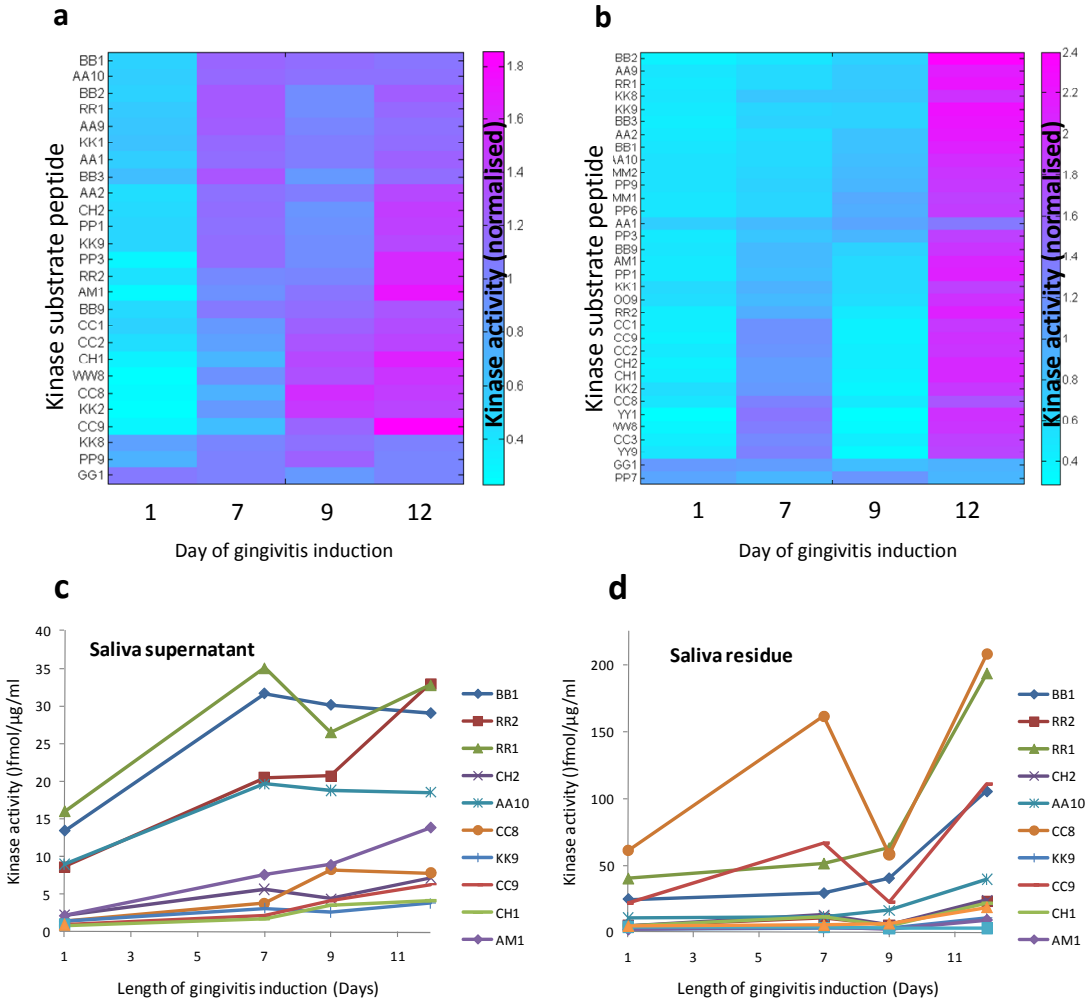


Figure 5.13 Heatmap of KAYAK activities in saliva (a) supernatant and (b) residue for day 1, 7, 9, 12 of gingivitis induction. Analysis by Pearson correlation hierarchical clustering was used to group similar responders together. The activity of each peptide was normalised to the sum of the kinase activities over all the days. Data are from patient 1 (water mouthwash).(c) and (d): Examples of kinase activity profiles in saliva supernatant and residue respectively.

In the saliva residue a large increase in kinase activity from day 1 to day 7 was observed, particularly for PKA and kinases targeting the basophilic substrate peptides (AA and BB respectively). PKC activity (CC substrate peptides) increased during days 7 to 9 of gingivitis induction. In the saliva residue, the kinase activities were generally higher than those in the supernatant (Figure 5.13c and d) and large increases in kinase activities were observed on the last day (day 12) compared to the supernatant, where a more gradual increase in kinase activity was observed.

The observed increase in kinase activity with gingivitis induction is likely to be as a result of activation of inflammatory signal pathways. Various kinases are thought to play a significant role in the activation and expression of inflammatory mediators, including the kinases IKK2, MAPK and PI3

kinase which can regulate inflammation through the activation of pro-inflammatory transcription factors such as NF-κB (nuclear factor κB) and AP-1 (activating protein-1). *Porphyromonas gingivalis* is known to activate the proinflammatory transcription factor NK-κB pathway, partly through the PI3 kinase/AKT pathway[415]. This is known to be followed by increased expression of anti-apoptotic genes and in the later stages of infection by increased pro-apoptotic genes[415-417].

The origin of kinase activity in the saliva supernatant is likely to be due to lysed cells, either through apoptosis or necrosis. Increased inflammation can lead to increased apoptosis which may explain the increased kinase activity in the saliva supernatant. Increased kinase activity in the saliva residue may be due to the presence of greater numbers of neutrophils in gingival tissue as well as the 'sloughing off' of gingival cells due to increased inflammation.

Various kinases were observed in the saliva supernatant and residue. In order to determine putative kinases responsible for the kinase activities observed during the course of gingivitis induction, the kinase activity and kinase abundance profiles were compared. Unfortunately, good correlation was not obtained on account of very low spectral counts of the kinases (data not shown). For future studies, the low-abundance kinases may be enriched by using hexapeptide libraries on beads (Proteominer, Biorad)[107], through immunodepletion of highly abundant proteins to allow better quantitation, or via fractionation of the saliva. Alternatively, more accurate quantitation methods (e.g. using AQUA peptides for each of the kinases) could be performed.

This proof-of-concept work indicates that kinase activity may be useful in investigating the changes in kinase activities during the course of gingivitis, and illustrates the potential of KAYAK to be performed in body fluids. For future studies it would be informative to separate the bacterial and human cells in saliva prior to cell lysis to enable the concentrations of human and bacterial proteins to be determined separately. Also, kinase activity profiling of gingival cells (cell culture) infected with putative periodontal pathogens could help elucidate the effect of the pathogen on the signalling pathways in the gingival cells.

5.4 Conclusions

The discovery of accurate biomarkers linked to the severity of gingivitis and periodontitis has important implications for patient care and management of the disease, as well as for the evaluation of different treatments and oral care hygiene. Various proof-of-concept techniques for the discovery of gingivitis biomarkers are reported here: saliva supernatant and saliva residue were evaluated using proteomics and the protein abundance of both the human and bacterial proteins were considered. Kinase activity profiling was performed on both the saliva supernatant and residue.

In the analysis of the first set of clinical samples from an internal Procter & Gamble clinical trial, the saliva samples from the healthy and gingivitis subjects were pooled, the saliva was fractionated into

10 gel pieces and analysed by GeLC-MSMS as well as being analysed individually by in-solution digestion and LC-MSMS. A large degree of individual variation was found and no significant changes in protein abundance in saliva between control and gingivitis subjects were clearly identified. In the induced gingivitis longitudinal clinical trial performed at the Eastman Dental Institute, inter-person variability was reduced and a time-profile of the changes in protein abundance in both the saliva supernatant and residue was obtained. On account of the large number of samples, fractionation was not feasible in terms of instrument time and analysis was performed using in solution LC-MSMS.

Various proteins previously reported to increase with gingivitis/periodontitis were also observed to increase with the induction of gingivitis; these proteins included MMP9, LDH, IL1F5, IL1F6, various members of S100, cystatin and cathepsin protein families, as well as various bacteria including *Fusobacterium nucleatum* and *Porphyromonas gingivalis*. Various new potential salivary biomarkers that merit further investigation include CRISP3, CALU and ERO1L proteins in the saliva supernatant, and LTA4H and PI3 proteins and *Haemophilus influenza* and *H. ducreyi* in the saliva residue.

Ideally, measuring the abundance of numerous human and bacterial biomarkers would allow for the accurate diagnosis and assessment of the severity of gingivitis. Utilising both host-response and pathogen biomarkers is likely to lead to the most accurate predictions of disease state and to allow the effectiveness of new drugs/oral care additives to be evaluated. Although not all the data from the induced gingivitis study could be included due to confidentiality, the analysis of one patient on a water treatment demonstrates the viability of this technique. These results suggest that a 2-hour mass spectrometric analysis of one sample can provide adequate information (around 600-1000 saliva proteins) for protein profiling of human and bacterial proteins.

To the author's knowledge, this is the first reported comparison study of induced gingivitis in humans using mass spectrometry based proteomics, as well as the first time that the saliva residue and bacteria have been assessed for biomarker discovery using proteomics. Kinase profiling of saliva in health and disease has not previously been attempted and may offer a promising alternative technique for biomarker discovery in saliva. These novel techniques lay the foundation for further development of diagnostic and prognostic tools and may provide fundamental insights into the cellular dynamics of inflammation and the complex host-pathogen interactions present during gingivitis.

6 Conclusion

6.1 Summary

The specific aims of this PhD project were the application of mass spectrometry technology to:

a) investigate the salivary proteome and develop a workflow for salivary biomarker discovery, and

b) optimise and develop a functional protein assay for measuring multiple kinase activities for kinome profiling.

With respect to the first aim, a number of methods were utilised to quantify various aspects of protein abundance within different portions of saliva. A saliva sample from a single individual was analysed using GeLC-MSMS proteomics. The saliva was split into the soluble portion, termed supernatant, and the insoluble portion, termed residue; and the two fractions were analysed for both human and bacterial proteins. A total of 834 and 1,426 human proteins in the saliva supernatant and residue respectively were identified. A large degree of overlap was observed between the saliva supernatant and residue, indicating that a significant proportion of the salivary supernatant proteins may originate from lysed cells.

By further examining the origins of the identified proteins (human or bacteria) revealed surprising numbers of bacterial species present in salivary fluids. Individual saliva samples from six patients were collected during a two week period of gingivitis induction in which five saliva samples were taken for each subject. Of the 2,654 proteins identified in the saliva residue, 1,378 were of human origin and 1,274 were from bacteria. Proportionally, the number of proteins of bacterial origin identified in saliva from a single healthy individual (9%) was almost five times lower than the number identified in six individuals in whom gingivitis was being induced. In the analysis of the saliva residue from six patients taken over a course of two weeks almost half of all the proteins identified in the saliva residue were from bacteria (48%). The high proportion of bacterial proteins in the saliva residue is likely due the effects of gingivitis inductions as well as the greater diversity of bacterial species between individual subjects. In this analysis, a total of 198 bacterial species were identified; this value is much higher than values previously reported by Xie, *et al.* (2008) in which 34 bacterial species were identified[211]. Proteomics research has, up until now, largely ignored the potential clinical and basic-science benefits of studying saliva residue and the bacterial proteome using proteomic tools. This study highlights the potential of using the saliva residue as a sample for biomarker studies and identifies that the bacterial component is a significant contributor to the proteins identified in the saliva residue. This indicates that future saliva studies may benefit from considering both bacterial and human proteome contributions in disease and biomarkers.

The presence of protein complexes in saliva was also investigated to gain insight into intracellular protein interactions. Following successful proof-of-concept work to profile protein complexes in whole cell lysate using BN PAGE and LC-MSMS[278], this technique was applied to investigate the presence of protein complexes in saliva. The presence of two main protein complexes were observed: one comprising predominately immunoglobulins, amylase and the kinase RIPK4, and the second containing mainly enzymes including LPO, MDH2, GADPH, CAT, ALDOA, CA6, PP1B and LTF. The technique of protein correlation profiling using BN PAGE was extended to measure the relative protein turnover of proteasomal proteins present in different proteasomal subunits using a modified SILAC experiment. The result of the extended BN PAGE SILAC experiment showed that the formation of the BP1 complex was faster than that of the 26S proteasome. This supports the previous finding that PSMD5 is a molecular chaperone involved in the formation of the BP1 complex (Appendix A.1)[278]. The success in identifying such protein complexes validates the BN PAGE protein correlation profiling technique, and demonstrates that this approach can be useful for investigating the protein composition of complexes and native protein-protein interactions. Furthermore, estimation of specific protein/protein-complex turnover rates in conjunction with isotopic labelling may be performed utilising BN PAGE protein correlation profiling.

To address the second major aim of this PhD project, a mass spectrometry based assay that measured multiple kinase activities analysis (KAYAK) was further developed and optimised. A new set of substrate peptides for the assay was investigated and the use of KAYAK was extended beyond that of human cell lysate to applications on various organisms including mouse, *Drosophila* and yeast. Significant kinase activity was also observed in both the saliva supernatant and residue using KAYAK. The success in applying KAYAK to a range of organisms and body fluid types demonstrates the tractability of the KAYAK method. The ease of application of KAYAK is significant for future applications of the method; production of test-kits for disease and cancer diagnosis, or a high-throughput 96-well plate assay for high throughput drug and kinase inhibitor screening are two possible applications of KAYAK. It is important to note that protein phosphorylation is not the only disease-relevant form of protein modification. Hundreds of different intracellular protein modifications exist, and a long term goal of the KAYAK method is to modify the assay to profile other types of activities, such as phosphatases and proteases.

To combine traditional proteomic analysis with this new KAYAK technique, in-gel kinase activity/kinase abundance correlation profiling using BN PAGE was performed. Kinase activity was observed following BN PAGE and various known kinase-substrate pairs were identified. Unfortunately, this procedure suffered due to the poor resolution of the BN PAGE but was useful for samples of lower complexity. For global proteome samples an alternative approach using AEX protein separation kinase activity/kinase abundance correlation profiling was used in this study. Using AEX separation the protein fractions were analysed using KAYAK to obtain kinase activity profiles, and by shotgun proteomics to obtain kinase abundance profiles. The kinase activity and kinase abundance profiles were correlated to match up substrate peptides and responsible kinases. Candidate kinases for

a number of documented hyperphosphorylated mitotic proteins from a recent phosphoproteomics study[176] without an identified cognate kinase were identified using this correlation profiling approach. After identification of potential kinase-substrate pairs using correlation profiling, pairs were validated using several different techniques. Putative kinase-substrate pairs were validated using *in vitro* kinase assays, overexpression of recombinant kinase, and siRNA knock down of the endogenous kinases. PAK3 was identified as the most likely candidate kinase for phosphorylation of the peptide substrate HFRNVHsEDFENR derived from the protein ADNP. The proposed phosphorylation of ADNP by PAK3 is of interest since PAK3 is thought to have some effect in neuronal cells and ADNP is known to be critical for brain and neuron development and this work could offer a link between these two proteins. PLK1 is the candidate kinase for the phosphorylation of the peptide substrates DDFKLNSsIVEPK, IEKLNSsLHFLQQ and AQHPDYsFGELSR that are derived from the proteins NFKB2, CAMSAP1L1 and PB1 respectively. PLK1 is highly upregulated in mitosis but has not previously been associated with the phosphorylation of NFKB2, CAMSAP1L1 or PB1. The putative phosphorylation of NFKB2 could be an important finding, since NFKB2 is a key transcription factor and very little is known about the transcriptional control of the cell cycle. This study has shown that by starting with a 'hit' from a phosphoproteomics experiment, kinase activity/kinase abundance correlation profiling can lead to the identification of putative kinases responsible for the protein phosphorylation.

As a way of applying these aforementioned techniques for potential gingivitis biomarker discovery, saliva samples from an induced gingivitis trial were analysed using proteomics to examine protein abundance and using KAYAK to obtain kinase activity profiles. The changes in the abundance of human and bacterial proteins in both the saliva supernatant and residue were profiled using spectral counting label-free proteomics. The kinase activities in both the supernatant and residue were determined using KAYAK. This work is ongoing, however preliminary data is promising. Several protein abundances and activities show a clear trend after the induction of gingivitis. Specific examples include the induction of MMP9, LDH, IL1F5, IL1F6, various members of S100, cystatin and cathepsin protein families, which have all been previously documented as induced or increased upon gingivitis/periodontitis. Similarly, various bacteria documented to increase in gingivitis, including *Fusobacterium nucleatum* and *Porphyromonas gingivalis*, were also increased in this study. Although the statistical power of this study was low (one patient) and clearly needs to be performed with a greater number of subjects, the confirmation of similar trends observed in the literature helps to validate this method and offers the opportunity for new biomarkers to be identified. In particular, CRISP3, CALU and ERO1L proteins in the saliva supernatant, and LTA4H and PI3 proteins and the species *Haemophilus influenza* and *H. ducreyi* in the saliva residue show good positive correlation with gingivitis induction and merit further investigation as potential new gingivitis biomarkers.

To discover not only protein abundance biomarkers, but potential protein activity biomarkers for gingivitis, a KAYAK assay was performed using these same saliva samples. A general increase in kinase activity was observed in both the saliva supernatant and residue with the induction of

gingivitis. This may be explained by the increased inflammation gingivitis causes as it progresses. Inflammation leads to an influx of leukocytes into gingival tissues, an activation of inflammatory cell signalling pathways, a release of inflammatory cytokines, and results in increased apoptosis/lysis. The increase in kinase activity in the saliva supernatant may be explained by this increase in gingival cell lysis, whilst the increase in kinase activity in the saliva residue is likely to be due to greater activation of inflammatory cell signalling pathways. We would expect different signalling pathways to be regulated differently during gingivitis induction, which is consistent with the varying kinase substrate activity profiles seen in the results presented in this thesis. This proof-of-concept work indicates that kinase activity profiling may be useful to monitor the extent of inflammation with gingivitis induction and could be useful in other inflammatory diseases like atherosclerosis, arthritis and autoimmune diseases.

It is important to note that the saliva residue contains many different cell types and is not a homogeneous mixture. Future profiling of the saliva residue would benefit from characterisation and separation of the human and bacterial cells to enable more accurate quantitation of changes in protein abundance against a background of human and bacterial proteins, respectively. With kinase activity profiling, it is especially important to determine an accurate human protein amount, since kinase activity originates predominately from human cells. It would also be of interest to profile the different types and proportions of human cells comprising the residue, for example leukocytes and gingival cells.

This PhD project has involved mass spectrometry to study a variety of systems using both ‘classical proteomics’ to quantitate protein abundance and ‘functional proteomics’ to profile kinase activities. Although the project has been based around gingivitis biomarker discovery in saliva, a wide variety of projects have evolved, merged and mixed with this original project aim. The merging of the KAYAK project with saliva biomarker discovery not only shows that KAYAK can be performed in body fluids and diagnostic information can be obtained on the state of inflammation, but it also offers huge potential for the functional protein profiling in saliva for other enzyme classes since the KAYAK strategy can be adapted to other enzyme classes. By utilising this project as a foundation, a long-term goal is the development of a ‘gingivitis spit test’. Further development of a gingivitis spit test would benefit greatly from the expansion of the KAYAK method to a protease activity assay, which is the aim of the next stage of the KAYAK project.

Many diseases, not only gingivitis and periodontitis, involve a complex interplay between bacterial and host responses. Therefore, to obtain a full picture of the diseases, it is useful to consider both the human and bacterial protein changes, along with protein activity measurements. The combination of both a functional protein assay and proteomic data is a unique way to search for biomarkers and offers great potential for future biomarker studies, as well as affords greater accuracy and specificity in biomarker analysis and diagnostics.

6.2 Concluding remarks

My time was divided between two main projects: a) investigating the salivary proteome and b) the development and optimisation of a multiplexed kinase assay using mass spectrometry: Kinase ActivitY Assay for Kinome profiling (KAYAK). Whilst these two seemingly disparate projects appear to have little in common, each of the projects has benefited from the other: the application of KAYAK to saliva has shown the presence of significant kinase activity in saliva, which has been previously unseen. This expands the future potential use of KAYAK on body fluids as well as showing the application of a new technique that can be applied to salivary biomarker discovery.

6.2.1 A short chronology of my PhD

I have been extremely fortunate to have had the opportunity to be based in two different mass spectrometry research groups: Dr Perdita Barran's group in the Chemistry Department at The University of Edinburgh (UoE) and Prof. Steven Gygi's group in the Cell Biology Department at Harvard Medical School (HMS). The first two years of my PhD were spent in the Barran group at UoE and involved familiarising myself with various different mass spectrometry and HPLC equipment. I was able to gain experience using an Applied Biosystems MALDI Voyager, a Waters Ultima QTOF, a Bruker High Capacity Ion trap and a Bruker 12T FTICR mass spectrometer. HPLC systems used included a Waters CapLC, Dionex Ultimate, Eksigent nano, Proxeon's 'Easy' nanoLC and the Dionex 3000 system. My subsequent two years work in the Gygi lab balanced this useful experience in Edinburgh of 'hands-on' mass spectrometry with a greater focus on biological application. The instrument set up at HMS comprised predominately Thermo LTQ Orbitrap mass spectrometers and Agilent HPLC systems.

In the course of my first year at Edinburgh I experimented with analysing my own saliva on the MALDI Voyager and Waters QTOF and investigating different ways to prepare saliva. This was helped by a week's trip to Prof. Joe Loo's lab at UCLA (University of California, Los Angeles) in December 2006 where they were working on the salivary proteome. I compared the saliva from several healthy people qualitatively using SDS-PAGE gels, and the amount of variation between people was clearly evident. During my second year at Edinburgh, I used predominately the Bruker HCT and FTICR mass spectrometer and focused on label-free quantitation to establish a work flow for biomarker discovery. Since the aim was to develop a platform for the analysis of a large number of samples, the plan was to use intensity based label-free quantitation. I attempted a comparison of three label-free software packages to determine which one would be most appropriate for the available mass spectrometers. Near the end of my second year I obtained saliva samples from healthy people and gingivitis sufferers from an internal P&G clinical trial. Knowing that the bacteria played a significant role in gingivitis, I became interested in analysing both the supernatant and the pellet remaining after centrifugation (the residue) as I thought that the most dramatic changes might be observed in the pellet

which contained the majority of the bacterial cells. One of the problems in the analysis of the pellet was the 'contamination' of the bacteria with insoluble human proteins, such as mucin, and it was difficult to obtain adequate separation. Through collaboration with the Eastman Dental Institute in London, I was able to investigate the bacterial portion more thoroughly by analysing the bacterial pellicle obtained from an artificial mouth model of gingivitis. Saliva was also cultured to increase the proportion of bacterial proteins over the 'contaminating' human ones. These initial analyses of salivary bacteria led to the probable detection of archaea in saliva. Realising that the salivary proteome was too complex to be analysed quantitatively in a single run, and that intensity based label-free quantitation with fractionation was currently too complicated for the equipment and bioinformatics resources available, I then considered analysing only a fraction of the salivary proteome, the phosphoprotein portion of saliva.

I moved to HMS in September 2008 and spent the third and fourth year of my PhD in the Gygi laboratory, where my time was spent both on saliva and a functional protein assay to measure kinase activities. For saliva, the proteome of both the supernatant and residue was investigated using GeLC-MSMS on an LTQ-Orbitrap. I carried out a comparison of the healthy and gingivitis saliva from the P&G internal trial. For quantitation I used spectral counting which allowed the fractionation of the sample and the identification of more proteins, whilst also the ability to quantify them, rather than the ion intensities method previously used at UoE.

Alongside my work on saliva, I worked on the new Kinase ActivitY Assay for Kinome profiling (KAYAK) technique which was being developed in the Gygi laboratory. The original assay used a collection of 90 substrate peptides. My projects involved evaluating a new substrate set comprising 60 peptides and applying the assay to a variety of systems and biological problems. Kinase activities in 12 human cell lysates were profiled and many differences were observed, each having a unique kinase activity signature. Kinase activity profiles in mice, *Drosophila* and yeast were also profiled under various biological conditions.

As well as the 'standard' KAYAK reaction, one of my projects involved proof-of-concept work to determine whether in-gel KAYAK was feasible and whether kinase activity/kinase abundance correlation profiling could be performed using blue native gels. One of the first steps in evaluating the method was to determine whether protein complexes remained intact following blue native polyacrylamide electrophoresis. The profiles of the proteins in the proteasome complex were investigated and, as well as showing that the complex remained intact, the method provided confirmation of the presence of the BP1 complex[278]. However, one of the main problems with this correlation technique was the poor resolution of the blue native gels which limited applicability of the correlation profiling method. An alternative chromatographic separation technique, anion exchange chromatography, was used instead to search for kinase-substrate pairs. The candidate kinase-substrate pairs were evaluated using several molecular biology techniques including overexpression using Gateway cloning and siRNA knock down. A slight variation of the kinase activity/kinase abundance

correlation profiling method was investigated using mouse tissues where the kinase activities in nine mouse tissues were compared with the kinase abundance.

During my tenure in the Gygi lab, I received saliva samples from an induced gingivitis clinical trial at the Eastman Dental Institute (London, UK). Having gained experience with the kinase acitivity assay, KAYAK, I applied this assay to monitor the changes in kinase activities during the induction of gingivitis. The induced gingivitis saliva samples were also investigated using proteomics: both the saliva supernatant and residue being analysed using LC-MSMS. The protein abundance was calculated using spectral counting and both human and bacterial proteins investigated.

6.2.2 Collaborative work

One of the advantages of working in a research area which involves highly technical instrumentation, such as mass spectrometry, is that the facilities are very much in demand by research groups needing analysis of their samples. This stimulates various collaborations with more biologically orientated research groups. In both the Barran and Gygi lab, I was involved in a variety of collaborations; many such collaborations were closely related to my work while others were often very different.

During my time with the Barran laboratory, I was involved with a *Caenorhabditis elegans* proteomics project which was part of the SCIBS (Selective Chemical Intervention in Biological Systems Initiative) project (involving the Walkinshaw group at Edinburgh, the Page group at the Institute of Comparative Medicine at Glasgow and the Turner group at Manchester); a sputum proteomics project with Dr Robert Gray of the Western Hospital in Edinburgh; a urine proteomics exosome project with Dr James Dear at the Edinburgh Royal Infirmary, and a mass spectrometry cross-linking project with Dr Malcolm Lyon in Manchester.

During my time in the Gygi group, I was also involved with several collaborations. A collaboration with the Finley lab (Cell Biology, HMS) involving the proteasome evolved from my work studying complexes with BN PAGE[278]. Another collaboration involved investigating the pupylome (the proteome of PUPylated proteins where PUP is a ubiquitin-like protein in prokaryotes) with the Darwin group at NYU School of Medicine[418]. I was involved with various collaborations directly related to KAYAK: a collaboration with the Perrimon lab (Genetics, HMS) involving *Drosophila*, the Weinstock lab (Dana Farber Cancer Institute, Boston) on leukaemia, and the Flier lab (HMS) on mice treated with fibroblast growth factor. I was also involved in a collaboration with Merck Research laboratories to develop an assay for atypical PKCs along with a post-doctoral fellow in the Gygi lab, Dr Ryan Kunz.

6.2.3 Final remarks

The original objective on starting this PhD was to find biomarkers for gingivitis. Whilst the 'perfect solution' of a gingivitis assay has not yet been achieved, the skills developed along the way may form the foundation for future development of such tools. This work may allow for methods which move beyond the traditional large scale proteomics 'healthy' versus 'diseased' comparison. Development of the functional protein assay for kinase activities using mass spectrometry (KAYAK) uses a strategy that can be applied to any enzyme class. The development of an assay for protease activity would offer huge diagnostic potential for gingivitis because saliva is highly abundant in such proteases. Protease activity can lead to receding gums, one of the symptoms of gingivitis, and quantifying this activity may be an efficient method of measuring the severity of the condition. Future work will concentrate on developing such a diagnostic tool. While the projects pursued in the course of this PhD may at first glance seem relatively unrelated, taken together, they add novel and fundamental pieces to the puzzle of the eventual development of sensitive and specific body fluid assays. In particular, a saliva based 'gingivitis assay' may benefit from the techniques presented here, and potential new biomarkers for the progression of this disease.

Whilst gingivitis does not appear at first sight to be a very exciting disease in comparison to diseases with higher mortality or morbidity, gingivitis affects a very large proportion of the population. The majority of treatments and investment are targeted towards reparative dentistry, rather than preventative measures, predominately due to fiscal incentives and the dental healthcare system. With good biomarkers and a better understanding of gingivitis and its progression to periodontitis, it is hoped that in the future dentistry products and care will focus more on preventative measures rather than expensive reparative ones. As with many medical treatments, early diagnosis and investment at the start should reduce the need for expensive late-stage care. It is likely that the oral microbiome plays a key role in the development of gingivitis and periodontitis. Once the key bacteria for a healthy mouth flora are determined, it may be envisioned that following an oral cleansing with chlorhexidine or a broad spectrum antibiotic, the oral microbiota could be repopulated with 'healthy bacteria' either in the form of a 'probiotic' drink or by direct coating of the mouth surfaces.

A note on the 'evolution' of the induced gingivitis project: during my 'Year In Industry' at P&G's Brussels Innovation Centre in 2004/05 I had a keen interest in undertaking a PhD which had the potential for translation into practical useful applications. After P&G kindly offered to support my PhD there were numerous ideas for projects: from the development of an 'active' to put into detergents to kill bed bugs, to a 'probiotic' drink' containing certain bacteria that could be given to children to reduce their risk of developing allergies. Having no real understanding of what a biomarker was, a telephone call with P&G suggesting a project in oral biomarkers led me to believe they desired something similar to the 'blue tablet' given by the dentist to stain the plaque on your teeth blue (which you then spend time trying to remove by brushing your teeth very thoroughly). However, a better understanding of the very wide uses of biomarkers made the project very appealing and has

given me exposure to a wide variety of fields. My original degree from the University of Edinburgh in 'Chemistry with Industrial Experience' included very little biology, which I had previously only studied to AS level at secondary school. This PhD project has stimulated me to study independently relevant areas of the subject and led to appreciate the application of chemical methodology to biological problems. My future contribution to biological research is strengthened, I feel, by my knowledge of chemistry. It can be strengthened further by combining this knowledge with the acquisition of highly relevant skills in computer programming. This is my envisaged way forward.

7 References

1. Hood, L., Heath, J.R., Phelps, M.E. & Lin, B. Systems biology and new technologies enable predictive and preventative medicine. *Science* **306**, 640-643 (2004).
2. Goldacre, B. Bad science. (Fourth Estate, London; 2008).
3. Weston, A.D. & Hood, L. Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine. *J Proteome Res* **3**, 179-196 (2004).
4. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat Biotechnol* **26**, 1135-1145 (2008).
5. Hood, L., Rowen, L., Galas, D.J. & Aitchison, J.D. Systems biology at the Institute for Systems Biology. *Brief Funct Genomic Proteomic* **7**, 239-248 (2008).
6. Petricoin, E.F., Zoon, K.C., Kohn, E.C., Barrett, J.C. & Liotta, L.A. Clinical proteomics: translating benchside promise into bedside reality. *Nat Rev Drug Discov* **1**, 683-695 (2002).
7. Brent, R. Genomic biology. *Cell* **100**, 169-183 (2000).
8. Rifai, N., Gillette, M.A. & Carr, S.A. Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nat Biotechnol* **24**, 971-983 (2006).
9. Godovac-Zimmermann, J. & Brown, L.R. Perspectives for mass spectrometry and functional proteomics. *Mass Spectrom Rev* **20**, 1-57 (2001).
10. Sleno, L. & Emili, A. Proteomic methods for drug target discovery. *Curr Opin Chem Biol* **12**, 46-54 (2008).
11. Frank, R. & Hargreaves, R. Clinical biomarkers in drug discovery and development. *Nat Rev Drug Discov* **2**, 566-580 (2003).
12. Hu, S., Loo, J.A. & Wong, D.T. Human saliva proteome analysis and disease biomarker discovery. *Expert Rev Proteomics* **4**, 531-538 (2007).
13. Kingsmore, S.F. Multiplexed protein measurement: technologies and applications of protein and antibody arrays. *Nat Rev Drug Discov* **5**, 310-320 (2006).
14. Wasinger, V.C. et al. Progress with gene-product mapping of the Mollicutes: Mycoplasma genitalium. *Electrophoresis* **16**, 1090-1094 (1995).
15. Wilkins, M.R. et al. From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Biotechnology (N Y)* **14**, 61-65 (1996).
16. Goodlett, D.R. & Yi, E.C. Proteomics without polyacrylamide: qualitative and quantitative uses of tandem mass spectrometry in proteome analysis. *Funct Integr Genomics* **2**, 138-153 (2002).
17. Husi, H. & Lisa, A.N. in International Review of Neurobiology, Vol. Volume 61 49-77 (Academic Press, 2004).
18. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198-207 (2003).
19. Zhang, H., Yan, W. & Aebersold, R. Chemical probes and tandem mass spectrometry: a strategy for the quantitative analysis of proteomes and subproteomes. *Curr Opin Chem Biol* **8**, 66-75 (2004).
20. Foster, L.J. et al. A mammalian organelle map by protein correlation profiling. *Cell.* **125**, 187-199 (2006).
21. Murray, C.I., Barrett, M. & Van Eyk, J.E. Assessment of ProteoExtract subcellular fractionation kit reveals limited and incomplete enrichment of nuclear subproteome from frozen liver and heart tissue. *Proteomics* **9**, 3934-3938 (2009).
22. Steen, H. & Mann, M. The ABC's (and XYZ's) of peptide sequencing. *Nat. Rev. Mol. Cell. Biol.* **5**, 699-711 (2004).
23. Yates, J.R., 3rd, Speicher, S., Griffin, P.R. & Hunkapiller, T. Peptide mass maps: a highly informative approach to protein identification. *Anal Biochem* **214**, 397-408 (1993).
24. Pappin, D.J., Hojrup, P. & Bleasby, A.J. Rapid identification of proteins by peptide-mass fingerprinting. *Curr. Biol.* **3**, 327-332 (1993).
25. Aebersold, R. & Goodlett, D.R. Mass Spectrometry in Proteomics. *Chem. Rev.* **101**, 269-296 (2001).
26. Roepstorff, P. & Fohlman, J. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed Mass Spectrom* **11**, 601 (1984).

27. Biemann, K. Mass spectrometry of peptides and proteins. *Annu Rev Biochem* **61**, 977-1010 (1992).
28. Faull, P. Exploring Gas-Phase Protein Conformations by Ion Mobility-Mass Spectrometry. *PhD Thesis The University of Edinburgh* (2009).
29. Dancik, V., Addona, T.A., Clauser, K.R., Vath, J.E. & Pevzner, P.A. De Novo Peptide Sequencing via Tandem Mass Spectrometry. *Journal of Computational Biology* **6**, 327-342 (1999).
30. Estabrook, G.F., McMorris, F.R. & Meacham, C.A. Comparison of Undirected Phylogenetic Trees Based on Subtrees of Four Evolutionary Units. *Systematic Zoology* **34**, 193-200 (1985).
31. Frank, A. & Pevzner, P. PepNovo: De Novo Peptide Sequencing via Probabilistic Network Modeling. *Anal. Chem.* **77**, 964-973 (2005).
32. MacCoss, M.J. & Matthews, D.E. Quantitative MS for proteomics: teaching a new dog old tricks. *Anal Chem* **77**, 294A-302A (2005).
33. Mueller, L.N., Brusniak, M.-Y., Mani, D.R. & Aebersold, R. An Assessment of Software Solutions for the Analysis of Mass Spectrometry Based Quantitative Proteomics Data. *J. Proteome Res.* **7**, 51-61 (2008).
34. Yan, W. & Chen, S.S. Mass spectrometry-based quantitative proteomic profiling. *Brief Funct Genomic Proteomic* **4**, 27-38 (2005).
35. Zhong, H., Marcus, S.L. & Li, L. Two-dimensional mass spectra generated from the analysis of 15N-labeled and unlabeled peptides for efficient protein identification and de novo peptide sequencing. *J Proteome Res* **3**, 1155-1163 (2004).
36. Bondarenko, P.V., Chelius, D. & Shaler, T.A. Identification and relative quantitation of protein mixtures by enzymatic digestion followed by capillary reversed-phase liquid chromatography-tandem mass spectrometry. *Anal Chem* **74**, 4741-4749 (2002).
37. Blondeau, F. et al. Tandem MS analysis of brain clathrin-coated vesicles reveals their critical involvement in synaptic vesicle recycling. *Proc Natl Acad Sci U S A* **101**, 3833-3838 (2004).
38. Old, W.M. et al. Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol Cell Proteomics* **4**, 1487-1502 (2005).
39. Gygi, S.P. et al. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **17**, 994-999 (1999).
40. Oda, Y., Huang, K., Cross, F.R., Cowburn, D. & Chait, B.T. Accurate quantitation of protein expression and site-specific phosphorylation. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 6591-6596 (1999).
41. André P. de Leenheer, L.M.T. Applications of isotope dilution-mass spectrometry in clinical chemistry, pharmacokinetics, and toxicology. *Mass Spectrometry Reviews* **11**, 249-307 (1992).
42. Gerber, S.A., Rush, J., Stemman, O., Kirschner, M.W. & Gygi, S.P. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 6940-6945 (2003).
43. Ross, P.L. et al. Multiplexed Protein Quantitation in Saccharomyces cerevisiae Using Amine-reactive Isobaric Tagging Reagents. *Mol Cell Proteomics* **3**, 1154-1169 (2004).
44. Gygi, S.P., Rist, B., Griffin, T.J., Eng, J. & Aebersold, R. Proteome analysis of low-abundance proteins using multidimensional chromatography and isotope-coded affinity tags. *J. Proteome. Res.* **1**, 47-54 (2002).
45. Eng, J.K., McCormack, A.L. & Yates, J.R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry* **5**, 976-989 (1994).
46. Ross, P.L. et al. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics.* **3**, 1154-1169 (2004).
47. Choe, L. et al. 8-plex quantitation of changes in cerebrospinal fluid protein expression in subjects undergoing intravenous immunoglobulin treatment for Alzheimer's disease. *Proteomics* **7**, 3651-3660 (2007).
48. Fenselau, C. A review of quantitative methods for proteomic studies. *J Chromatogr B Analyt Technol Biomed Life Sci* **855**, 14-20 (2007).
49. Yao, X., Afonso, C. & Fenselau, C. Dissection of proteolytic 18O labeling: endoprotease-catalyzed 16O-to-18O exchange of truncated peptide substrates. *J. Proteome. Res.* **2**, 147-152 (2003).

50. Washburn, M.P., Ulaszek, R., Deciu, C., Schieltz, D.M. & Yates, J.R., 3rd Analysis of quantitative proteomic data generated via multidimensional protein identification technology. *Anal Chem* **74**, 1650-1657 (2002).
51. MacCoss, M.J., Wu, C.C., Liu, H., Sadygov, R. & Yates, J.R., 3rd A correlation algorithm for the automated quantitative analysis of shotgun proteomics data. *Anal Chem* **75**, 6912-6921 (2003).
52. Engelsberger, W.R., Erban, A., Kopka, J. & Schulze, W.X. Metabolic labeling of plant cell cultures with K(15)NO3 as a tool for quantitative analysis of proteins and metabolites. *Plant Methods* **2**, 14 (2006).
53. Conrads, T.P. et al. Quantitative analysis of bacterial and mammalian proteomes using a combination of cysteine affinity tags and 15N-metabolic labeling. *Anal. Chem.* **73**, 2132-2139 (2001).
54. Krijgsveld, J. et al. Metabolic labeling of C. elegans and D. melanogaster for quantitative proteomics. *Nat Biotechnol* **21**, 927-931 (2003).
55. Wu, C.C., MacCoss, M.J., Howell, K.E., Matthews, D.E. & Yates, J.R.r. Metabolic labeling of mammalian organisms with stable isotopes for quantitative proteomic analysis. *Anal. Chem.* **76**, 4951-4959 (2004).
56. Ong, S.-E. et al. Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics. *Mol Cell Proteomics* **1**, 376-386 (2002).
57. Ong, S.E., Foster, L.J. & Mann, M. Mass spectrometric-based approaches in quantitative proteomics. *Methods* **29**, 124-130 (2003).
58. Ong, S.E., Kratchmarova, I. & Mann, M. Properties of 13C-substituted arginine in stable isotope labeling by amino acids in cell culture (SILAC). *J. Proteome. Res.* **2**, 173-181 (2003).
59. Engen, J.R., Bradbury, E.M. & Chen, X. Using stable-isotope-labeled proteins for hydrogen exchange studies in complex mixtures. *Anal Chem* **74**, 1680-1686 (2002).
60. Chelius, D. & Bondarenko, P.V. Quantitative Profiling of Proteins in Complex Mixtures Using Liquid Chromatography and Mass Spectrometry. *J. Proteome Res.* **1**, 317-323 (2002).
61. Wang, H. & Hanash, S. Multi-dimensional liquid phase based separations in proteomics. *J. Chromatogr. B.* **787**, 11-18 (2003).
62. Nesvizhskii, A.I., Vitek, O. & Aebersold, R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods.* **4**, 787-797 (2007).
63. America, A.H. et al. Alignment and statistical difference analysis of complex peptide data sets generated by multidimensional LC-MS. *Proteomics* **6**, 641-653 (2006).
64. Lilley, K.S. & Dupree, P. Methods of quantitative proteomics and their application to plant organelle characterization. *J. Exp. Bot.* **57**, 1493-1499 (2006).
65. Bantscheff, M., Schirle, M., Sweetman, G., Rick, J. & Kuster, B. Quantitative mass spectrometry in proteomics: a critical review. *Anal Bioanal Chem* **389**, 1017-1031 (2007).
66. America, A.H. & Cordewener, J.H. Comparative LC-MS: a landscape of peaks and valleys. *Proteomics* **8**, 731-749 (2008).
67. Washburn, M.P., Wolters, D. & Yates, J.R. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotech* **19**, 242-247 (2001).
68. Hrapovic, S., Liu, Y., Male, K.B. & Luong, J.H.T. Electrochemical Biosensing Platforms Using Platinum Nanoparticles and Carbon Nanotubes. *Anal. Chem.* **76**, 1083-1088 (2004).
69. Dong, M.Q. et al. Quantitative mass spectrometry identifies insulin signaling targets in C. elegans. *Science.* **317**, 660-663 (2007).
70. Zybailov, B., Coleman, M.K., Florens, L. & Washburn, M.P. Correlation of relative abundance ratios derived from peptide ion chromatograms and spectrum counting for quantitative proteomic analysis using stable isotope labeling. *Anal Chem* **77**, 6218-6224 (2005).
71. Florens, L. et al. Analyzing chromatin remodeling complexes using shotgun proteomics and normalized spectral abundance factors. *Methods* **40**, 303-311 (2006).
72. Mueller, L.N. et al. SuperHirn - a novel tool for high resolution LC-MS-based peptide/protein profiling. *Proteomics* **7**, 3470-3480 (2007).
73. Old, W.M. et al. Comparison of Label-free Methods for Quantifying Human Proteins by Shotgun Proteomics. *Mol Cell Proteomics* **4**, 1487-1502 (2005).
74. Pang, J.X., Ginanni, N., Dongre, A.R., Hefta, S.A. & Opitek, G.J. Biomarker discovery in urine by proteomics. *J Proteome Res* **1**, 161-169 (2002).

75. Rao, P.V. et al. Proteomic identification of salivary biomarkers of type-2 diabetes. *J Proteome Res* **8**, 239-245 (2009).
76. Pan, J., Chen, H.Q., Sun, Y.H., Zhang, J.H. & Luo, X.Y. Comparative proteomic analysis of non-small-cell lung cancer and normal controls using serum label-free quantitative shotgun technology. *Lung* **186**, 255-261 (2008).
77. Venable, J.D., Dong, M.-Q., Wohlschlegel, J., Dillin, A. & Yates, J.R.r. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat. Methods.* **1**, 39-45 (2004).
78. Ting, L.J. A quantitative proteomics investigation of cold adaptation in the marine bacterium *Sphingopyxis alaskensis*. *PhD dissertation, University of New South Wales* (2010).
79. Schulze, W.X. & Mann, M. A novel proteomic screen for peptide-protein interactions. *J Biol Chem* **279**, 10756-10764 (2004).
80. Han, D.K., Eng, J., Zhou, H. & Aebersold, R. Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. *Nat. Biotechnol.* **19**, 946-951 (2001).
81. Li, X.J., Zhang, H., Ranish, J.A. & Aebersold, R. Automated statistical analysis of protein abundance ratios from data generated by stable-isotope dilution and tandem mass spectrometry. *Anal Chem* **75**, 6648-6657 (2003).
82. Bouyssie, D. et al. Mascot file parsing and quantification (MFPaQ), a new software to parse, validate, and quantify proteomics data generated by ICAT and SILAC mass spectrometric analyses: application to the proteomics study of membrane proteins from primary human endothelial cells. *Mol Cell Proteomics* **6**, 1621-1637 (2007).
83. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367-1372 (2008).
84. Graumann, J. et al. SILAC-labeling and proteome quantitation of mouse embryonic stem cells to a depth of 5111 proteins. *Mol Cell Proteomics*, M700460-MCP700200 (2007).
85. Park, S.K., Venable, J.D., Xu, T. & Yates, J.R., 3rd A quantitative analysis software tool for mass spectrometry-based proteomics. *Nat Methods* **5**, 319-322 (2008).
86. Andreev, V.P. et al. New algorithm for 15N/14N quantitation with LC-ESI-MS using an LTQ-FT mass spectrometer. *J Proteome Res* **5**, 2039-2045 (2006).
87. Palagi, P.M. et al. MSight: an image analysis software for liquid chromatography-mass spectrometry. *Proteomics* **5**, 2381-2384 (2005).
88. Lu, P., Vogel, C., Wang, R., Yao, X. & Marcotte, E.M. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol* **25**, 117-124 (2007).
89. Leptos, K.C., Sarracino, D.A., Jaffe, J.D., Krastins, B. & Church, G.M. MapQuant: open-source software for large-scale protein quantification. *Proteomics* **6**, 1770-1782 (2006).
90. Jaffe, J.D. et al. PEPPeR, a platform for experimental proteomic pattern recognition. *Mol Cell Proteomics* **5**, 1927-1941 (2006).
91. Monroe, M.E., Shaw, J.L., Daly, D.S., Adkins, J.N. & Smith, R.D. MASIC: a software program for fast quantitation and flexible visualization of chromatographic profiles from detected LC-MS(/MS) features. *Comput Biol Chem* **32**, 215-217 (2008).
92. May, D. et al. A platform for accurate mass and time analyses of mass spectrometry data. *J Proteome Res* **6**, 2685-2694 (2007).
93. Li, X.J., Yi, E.C., Kemp, C.J., Zhang, H. & Aebersold, R. A software suite for the generation and comparison of peptide arrays from sets of data collected by liquid chromatography-mass spectrometry. *Mol Cell Proteomics* **4**, 1328-1340 (2005).
94. Halligan, B.D. et al. ZoomQuant: an application for the quantitation of stable isotope labeled peptides. *J Am Soc Mass Spectrom* **16**, 302-306 (2005).
95. Lin, W.T. et al. Multi-Q: a fully automated tool for multiplexed protein quantitation. *J Proteome Res* **5**, 2328-2338 (2006).
96. Keller, A., Eng, J., Zhang, N., Li, X.J. & Aebersold, R. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol Syst Biol* **1**, 2005 0017 (2005).
97. Shadforth, I.P., Dunkley, T.P., Lilley, K.S. & Bessant, C. i-Tracker: for quantitative proteomics using iTRAQ. *BMC Genomics* **6**, 145 (2005).
98. Smith CA, O.M.G., Want EJ, Qin C, Trauger SA, Brandon TR, Custodio DE, Abagyan R, Siuzdak G METLIN: a metabolite mass spectral database. *Ther Drug Monit* **28**, 402-406.

(2006).
99. Journet, A. & Ferro, M. The potentials of MS-based subproteomic approaches in medical science: the case of lysosomes and breast cancer. *Mass Spectrom Rev* **23**, 393-442 (2004).
100. Schmidt, A. & Aebersold, R. High-accuracy proteome maps of human body fluids. *Genome Biol* **7**, 242 (2006).
101. Beavis, R.C. Using the global proteome machine for protein identification. *Methods Mol Biol* **328**, 217-228 (2006).
102. The Global Proteome Machine www.thegpm.org 2006).
103. Desiere, F. et al. Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biology* **6**, R9 (2004).
104. http://www.peptideatlas.org Peptide Atlas.
105. States, D.J. et al. Challenges in deriving high-confidence protein identifications from data gathered by a HUPO plasma proteome collaborative study. *Nat Biotech* **24**, 333-338 (2006).
106. Adachi, J., Kumar, C., Zhang, Y., Olsen, J.V. & Mann, M. The human urinary proteome contains more than 1500 proteins, including a large proportion of membrane proteins. *Genome Biol* **7**, R80 (2006).
107. Bandhakavi, S., Stone, M.D., Onsongo, G., Van Riper, S.K. & Griffin, T.J. A dynamic range compression and three-dimensional peptide fractionation analysis platform expands proteome coverage and the diagnostic potential of whole saliva. *J Proteome Res* **8**, 5590-5600 (2009).
108. Thompson, E.J. & Keir, G. Laboratory investigation of cerebrospinal fluid proteins. *Ann Clin Biochem* **27 (Pt 5)**, 425-435 (1990).
109. Raymackers, J. et al. Identification of two-dimensionally separated human cerebrospinal fluid proteins by N-terminal sequencing, matrix-assisted laser desorption/ionization--mass spectrometry, nanoliquid chromatography-electrospray ionization-time of flight-mass spectrometry, and tandem mass spectrometry. *Electrophoresis* **21**, 2266-2283 (2000).
110. Sickmann, A. et al. Identification of proteins from human cerebrospinal fluid, separated by two-dimensional polyacrylamide gel electrophoresis. *Electrophoresis* **21**, 2721-2728 (2000).
111. Yuan, X., Russell, T., Wood, G. & Desiderio, D.M. Analysis of the human lumbar cerebrospinal fluid proteome. *Electrophoresis* **23**, 1185-1196 (2002).
112. Bergquist, J., Palmblad, M., Wetterhall, M., Hakansson, P. & Markides, K.E. Peptide mapping of proteins in human body fluids using electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. *Mass Spectrom Rev* **21**, 2-15 (2002).
113. de Souza, G.A., Godoy, L.M. & Mann, M. Identification of 491 proteins in the tear fluid proteome reveals a large number of proteases and protease inhibitors. *Genome Biol* **7**, R72 (2006).
114. Yamakawa, K., Yoshida, K., Nishikawa, H., Kato, T. & Iwamoto, T. Comparative analysis of interindividual variations in the seminal plasma proteome of fertile men with identification of potential markers for azoospermia in infertile patients. *J Androl* **28**, 858-865 (2007).
115. States, D.J. et al. Challenges in deriving high-confidence protein identifications from data gathered by a HUPO plasma proteome collaborative study. *Nat. Biotechnol.* **24**, 333-338 (2006).
116. Omenn, G.S. et al. Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. *Proteomics.* **5**, 3226-3245 (2005).
117. Zhang, H. et al. Mass spectrometric detection of tissue proteins in plasma. *Mol Cell Proteomics* **6**, 64-71 (2007).
118. Rehman, I. et al. Proteomic analysis of voided urine after prostatic massage from patients with prostate cancer: A pilot study. *Urology* **64**, 1238-1243 (2004).
119. Thongboonkerd, V. & Malasit, P. Renal and urinary proteomics: current applications and challenges. *Proteomics* **5**, 1033-1042 (2005).
120. Julio E. Celis, H.W.M.Ø. Bladder squamous cell carcinoma biomarkers derived from proteomics. *Electrophoresis* **21**, 2115-2121 (2000).
121. Theodorescu, D. et al. Discovery and validation of new protein biomarkers for urothelial cancer: a prospective analysis. *The Lancet Oncology* **7**, 230-240 (2006).
122. Nicholas, B. et al. Shotgun proteomic analysis of human-induced sputum. *Proteomics* **6**, 4390-4401 (2006).
123. Collins, M.O., Yu, L. & Choudhary, J.S. Analysis of protein phosphorylation on a proteome-scale. *Proteomics.* **7**, 2751-2768 (2007).

124. Joseph, E.I. et al. An integrated functional genomics and metabolomics approach for defining poor prognosis in human neuroendocrine cancers. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 9901-9906 (2005).
125. Schulenberg, B., Aggeler, R., Beechem, J.M., Capaldi, R.A. & Patton, W.F. Analysis of Steady-state Protein Phosphorylation in Mitochondria Using a Novel Fluorescent Phosphosensor Dye. *J. Biol. Chem.* **278**, 27251-27255 (2003).
126. Raggiaschi, R. et al. Detection of phosphorylation patterns in rat cortical neurons by combining phosphatase treatment and DIGE technology. *Proteomics* **6**, 748-756 (2006).
127. Yamagata, A. et al. Mapping of phosphorylated proteins on two-dimensional polyacrylamide gels using protein phosphatase. *Proteomics* **2**, 1267-1276 (2002).
128. Li, Y., Qi, D., Deng, C., Yang, P. & Zhang, X. Cerium ion-chelated magnetic silica microspheres for enrichment and direct determination of phosphopeptides by matrix-assisted laser desorption ionization mass spectrometry. *J Proteome Res* **7**, 1767-1777 (2008).
129. Rush, J. et al. Immunoaffinity profiling of tyrosine phosphorylation in cancer cells. *Nat Biotech* **23**, 94-101 (2005).
130. Steen, H., Kuster, B., Fernandez, M., Pandey, A. & Mann, M. Tyrosine Phosphorylation Mapping of the Epidermal Growth Factor Receptor Signaling Pathway. *J. Biol. Chem.* **277**, 1031-1039 (2002).
131. Judit, V.n., Sean, A.B., Scott, A.G. & Steven, P.G. Large-scale phosphorylation analysis of mouse liver. *Proceedings of the National Academy of Sciences* **104**, 1488-1493 (2007).
132. Trinidad, J.C., Specht, C.G., Thalhammer, A., Schoepfer, R. & Burlingame, A.L. Comprehensive Identification of Phosphorylation Sites in Postsynaptic Density Preparations. *Mol Cell Proteomics* **5**, 914-922 (2006).
133. Ficarro, S.B. et al. Phosphoproteome analysis by mass spectrometry and its application to Saccharomyces cerevisiae. *Nat Biotech* **20**, 301-305 (2002).
134. Larsen, M.R., Thingholm, T.E., Jensen, O.N., Roepstorff, P. & Jorgensen, T.J.D. Highly Selective Enrichment of Phosphorylated Peptides from Peptide Mixtures Using Titanium Dioxide Microcolumns. *Mol Cell Proteomics* **4**, 873-886 (2005).
135. Bodenmiller, B., Mueller, L.N., Mueller, M., Domon, B. & Aebersold, R. Reproducible isolation of distinct, overlapping segments of the phosphoproteome. *Nat Meth* **4**, 231-237 (2007).
136. Ytterberg, J.e.a. Isolation and Identification of Phosphopeptides by Combining Strong Cationic Exchange and Hydrophilic Interaction Chromatography. *Poster Number 523* **ASMS** (2007).
137. Collins, M.O. et al. Proteomic Analysis of in Vivo Phosphorylated Synaptic Proteins. *J. Biol. Chem.* **280**, 5972-5982 (2005).
138. Zappacosta, F. et al. Improved Sensitivity for Phosphopeptide Mapping Using Capillary Column HPLC and Microionspray Mass Spectrometry: Comparative Phosphorylation Site Mapping from Gel-Derived Proteins. *Anal. Chem.* **74**, 3221-3231 (2002).
139. Williamson, B.L., Marchese, J. & Morrice, N.A. Automated Identification and Quantification of Protein Phosphorylation Sites by LC/MS on a Hybrid Triple Quadrupole Linear Ion Trap Mass Spectrometer. *Mol Cell Proteomics* **5**, 337-346 (2006).
140. Steen, H., Kuster, B. & Mann, M. Quadrupole time-of-flight versus triple-quadrupole mass spectrometry for the determination of phosphopeptides by precursor ion scanning. *J Mass Spectrom* **36**, 782-790 (2001).
141. Beausoleil, S.A. et al. Large-scale characterization of HeLa cell nuclear phosphoproteins. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 12130-12135 (2004).
142. Cooper, H.J., Hakansson, K. & Marshall, A.G. The role of electron capture dissociation in biomolecular analysis. *Mass Spectrom Rev* **24**, 201-222 (2005).
143. Mikesh, L.M. et al. The utility of ETD mass spectrometry in proteomic analysis. *Biochimica et Biophysica Acta (BBA) - Proteins & Proteomics* **1764**, 1811-1822 (2006).
144. Syka, J.E.P., Coon, J.J., Schroeder, M.J., Shabanowitz, J. & Hunt, D.F. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 9528-9533 (2004).
145. Villanueva, J. et al. A sequence-specific exopeptidase activity test (SSEAT) for "functional" biomarker discovery. *Mol Cell Proteomics* **7**, 509-518 (2008).
146. Jessani, N. et al. A streamlined platform for high-content functional proteomics of primary human specimens. *Nat Methods* **2**, 691-697 (2005).

147. Adam, G.C., Sorensen, E.J. & Cravatt, B.F. Proteomic profiling of mechanistically distinct enzyme classes using a common chemotype. *Nat. Biotechnol.* **20**, 805-809 (2002).
148. Greenbaum, D. et al. Chemical approaches for functionally probing the proteome. *Mol Cell Proteomics* **1**, 60-68 (2002).
149. Kato, D. et al. Activity-based probes that target diverse cysteine protease families. *Nat Chem Biol* **1**, 33-38 (2005).
150. Kubota, K. et al. Sensitive multiplexed analysis of kinase activities and activity-based kinase identification. *Nat Biotechnol* **27**, 933-940 (2009).
151. Yu, Y. et al. A site-specific, multiplexed kinase activity assay using stable-isotope dilution and high-resolution mass spectrometry. *Proc Natl Acad Sci U S A* **106**, 11606-11611 (2009).
152. Manning, G., Whyte, D.B., Martinez, R., Hunter, T. & Sudarsanam, S. The protein kinase complement of the human genome. *Science.* **298**, 1912-1934 (2002).
153. Cohen, P. The role of protein phosphorylation in human health and disease. The Sir Hans Krebs Medal Lecture. *Eur J Biochem* **268**, 5001-5010 (2001).
154. Cozzone, A.J. Protein phosphorylation in prokaryotes. *Annu Rev Microbiol* **42**, 97-125 (1988).
155. Hanahan, D. & Weinberg, R.A. The hallmarks of cancer. *Cell* **100**, 57-70 (2000).
156. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061-1068 (2008).
157. Ren, R. Mechanisms of BCR-ABL in the pathogenesis of chronic myelogenous leukaemia. *Nat Rev Cancer* **5**, 172-183 (2005).
158. Yeatman, T.J. A renaissance for SRC. *Nat Rev Cancer* **4**, 470-480 (2004).
159. Irish, J.M. et al. Single cell profiling of potentiated phospho-protein networks in cancer cells. *Cell* **118**, 217-228 (2004).
160. Eglen, R.M. & Reisine, T. The current status of drug discovery against the human kinome. *Assay Drug Dev Technol* **7**, 22-43 (2009).
161. ter Haar, E. et al. Kinase chemogenomics: targeting the human kinome for target validation and drug discovery. *Mini Rev Med Chem* **4**, 235-253 (2004).
162. Kim, C., Xuong, N.H. & Taylor, S.S. Crystal structure of a complex between the catalytic and regulatory (RIalpha) subunits of PKA. *Science* **307**, 690-696 (2005).
163. Diks, S.H. et al. Kinome profiling for studying lipopolysaccharide signal transduction in human peripheral blood mononuclear cells. *J. Biol. Chem.* **279**, 49206-49213 (2004).
164. Houseman, B.T., Huh, J.H., Kron, S.J. & Mrksich, M. Peptide chips for the quantitative evaluation of protein kinase activity. *Nat. Biotechnol.* **20**, 270-274 (2002).
165. Janes, K.A. et al. A high-throughput quantitative multiplex kinase assay for monitoring information flow in signaling networks: application to sepsis-apoptosis. *Mol Cell Proteomics* **2**, 463-473 (2003).
166. Shults, M.D., Janes, K.A., Lauffenburger, D.A. & Imperiali, B. A multiplexed homogeneous fluorescence-based assay for protein kinase activity in cell lysates. *Nat. Methods.* **2**, 277-283 (2005).
167. Blaydes, J.P., Vojtesek, B., Bloomberg, G.B. & Hupp, T.R. The development and use of phospho-specific antibodies to study protein phosphorylation. *Methods Mol Biol* **99**, 177-189 (2000).
168. Yu, Y. et al. A site-specific, multiplexed kinase activity assay using stable-isotope dilution and high-resolution mass spectrometry. *PNAS.* **106**, 11606-11611 (2009).
169. Stegmeier, F., Warmuth, M., Sellers, W.R. & Dorsch, M. Targeted cancer therapies in the twenty-first century: lessons from imatinib. *Clin Pharmacol Ther* **87**, 543-552 (2010).
170. Klumpp, M. et al. Readout technologies for highly miniaturized kinase assays applicable to high-throughput screening in a 1536-well format. *J Biomol Screen* **11**, 617-633 (2006).
171. Whitney, J.A. Reference systems for kinase drug discovery: chemical genetic approaches to cell-based assays. *Assay Drug Dev Technol* **2**, 417-429 (2004).
172. Jia, Y., Quinn, C.M., Kwak, S. & Talanian, R.V. Current in vitro kinase assay technologies: the quest for a universal format. *Curr Drug Discov Technol* **5**, 59-69 (2008).
173. Mok, J., Im, H. & Snyder, M. Global identification of protein kinase substrates by protein microarray analysis. *Nat Protoc* **4**, 1820-1827 (2009).
174. Thiele, A., Zerweck, J. & Schutkowski, M. Peptide arrays for enzyme profiling. *Methods Mol Biol* **570**, 19-65 (2009).
175. Jalal, S. et al. Genome to kinome: species-specific peptide arrays for kinome analysis. *Sci*

Signal **2**, pl1 (2009).
176. Dephoure, N. et al. A quantitative atlas of mitotic phosphorylation. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 10762-10767 (2008).
177. Huttlin, E. A Tissue-Specific Atlas of Mouse Protein Phosphorylation and Expression. *Submitted* (2010).
178. Zaccolo, M. & Pozzan, T. Discrete microdomains with high concentration of cAMP in stimulated rat neonatal cardiac myocytes. *Science* **295**, 1711-1715 (2002).
179. Perillan, P.R., Chen, M., Potts, E.A. & Simard, J.M. Transforming growth factor-beta 1 regulates Kir2.3 inward rectifier K+ channels via phospholipase C and protein kinase C-delta in reactive astrocytes from adult rat brain. *J Biol Chem* **277**, 1974-1980 (2002).
180. Hirtz C, C.F., Centeno D, Egea JC, Rossignol M, Sommerer N, de Périère D Complexity of the human whole saliva proteome. *J Physiol Biochem.* **61**, 469-480 (2005).
181. Tabak, L.A. Structure and function of human salivary mucins. *Crit Rev Oral Biol Med* **1**, 229-234 (1990).
182. Humphrey, S.P. & Williamson, R.T. A review of saliva: normal composition, flow, and function. *J Prosthet Dent* **85**, 162-169 (2001).
183. Sreebny, L.M. & Valdini, A. Xerostomia. A neglected symptom. *Arch Intern Med* **147**, 1333-1337 (1987).
184. Buser, P. Slowly forgetting the Pavlovian adventure? *C R Biol* **329**, 398-405 (2006).
185. Edgar, W.M. Saliva and dental health. Clinical implications of saliva: report of a consensus meeting. *Br Dent J* **169**, 96-98 (1990).
186. Ishikawa, Y. et al. Water channels and zymogen granules in salivary glands. *J Pharmacol Sci* **100**, 495-512 (2006).
187. Gonzalez-Begne, M. et al. Proteomic analysis of human parotid gland exosomes by multidimensional protein identification technology (MudPIT). *J Proteome Res* **8**, 1304-1314 (2009).
188. Huang, A.Y., Castle, A.M., Hinton, B.T. & Castle, J.D. Resting (basal) secretion of proteins is provided by the minor regulated and constitutive-like pathways and not granule exocytosis in parotid acinar cells. *J Biol Chem* **276**, 22296-22306 (2001).
189. Wong, D.T. Salivary diagnostics powered by nanotechnologies, proteomics and genomics. *J Am Dent Assoc* **137**, 313-321 (2006).
190. Tabak, L.A., Levine, M.J., Mandel, I.D. & Ellison, S.A. Role of salivary mucins in the protection of the oral cavity. *J Oral Pathol* **11**, 1-17 (1982).
191. Hu S, D.P., Denny P, Xie Y, Loo JA, Wolinsky LE, Li Y, McBride J, Ogorzalek Loo RR, Navazesh M, Wong DT Differentially expressed protein markers in human submandibular and sublingual secretions. *Int J Oncol.* **25**, 1423-1430 (2004).
192. Guo, T. et al. Characterization of the Human Salivary Proteome by Capillary Isoelectric Focusing/Nanoreversed-Phase Liquid Chromatography Coupled with ESI-Tandem MS. *J. Proteome Res.* **5**, 1469-1478 (2006).
193. Wilmarth, P.A. et al. Two-Dimensional Liquid Chromatography Study of the Human Whole Saliva Proteome. *J. Proteome Res.* **3**, 1017-1023 (2004).
194. Huang, C.-M. Comparative proteomic analysis of human whole saliva. *Archives of Oral Biology* **49**, 951-962 (2004).
195. Hu S, D.P., Denny P, Xie Y, Loo JA, Wolinsky LE, Li Y, McBride J, Ogorzalek Loo RR, Navazesh M, Wong DT Differentially expressed protein markers in human submandibular and sublingual secretions. *Int J Oncol* **25**, 1423-1430 (2004).
196. Hardt, M. et al. Toward Defining the Human Parotid Gland Salivary Proteome and Peptidome: Identification and Characterization Using 2D SDS-PAGE, Ultrafiltration, HPLC, and Mass Spectrometry. *Biochemistry* **44**, 2885-2899 (2005).
197. Shen Hu, Y.X.P.R.R.R.O.L.Y.L.J.A.L.D.T.W. Large-scale identification of proteins in human salivary proteome by liquid chromatography/mass spectrometry and two-dimensional gel electrophoresis-mass spectrometry. *PROTEOMICS* **5**, 1714-1728 (2005).
198. Amado, F.M.L. et al. Analysis of the human saliva proteome. *Expert Review of Proteomics* **2**, 521-539 (2005).
199. Anke Walz, K.S., Andreas Wattenberg, Eva Hawranke, Helmut E Meyer, Gottfried Schmalz, Martin Blüggel, Stefan Ruhl Proteome analysis of glandular parotid and submandibular-sublingual saliva in comparison to whole human saliva by two-dimensional gel electrophoresis. *PROTEOMICS* **6**, 1631-1639 (2006).

200. Savitski, M.M., Nielsen, M.L. & Zubarev, R.A. ModifiComb, a New Proteomic Tool for Mapping Substoichiometric Post-translational Modifications, Finding Novel Types of Modifications, and Fingerprinting Complex Protein Mixtures. *Mol Cell Proteomics* **5**, 935-948 (2006).
201. Vitorino, R. et al. Two-dimensional electrophoresis study of in vitro pellicle formation and dental caries susceptibility. *European Journal of Oral Sciences* **114**, 147-153 (2006).
202. Ramachandran, P. et al. Identification of N-Linked Glycoproteins in Human Saliva by Glycoprotein Capture and Mass Spectrometry. *J. Proteome Res.* **5**, 1493-1503 (2006).
203. Beeley JA, K.K., Lamey PJ Two-dimensional electrophoresis of human parotid salivary proteins from normal and connective tissue disorder subjects using immobilised pH gradients. *Electrophoresis* **12**, 493-499 (1991).
204. Yao, Y., Berg, E.A., Costello, C.E., Troxler, R.F. & Oppenheim, F.G. Identification of Protein Components in Human Acquired Enamel Pellicle and Whole Saliva Using Novel Proteomics Approaches. *J. Biol. Chem.* **278**, 5300-5308 (2003).
205. Denny, P. et al. The proteomes of human parotid and submandibular/sublingual gland salivas collected as the ductal secretions. *J Proteome Res* **7**, 1994-2006 (2008).
206. Denny, P. et al. The Proteomes of Human Parotid and Submandibular/Sublingual Gland Salivas Collected as the Ductal Secretions. *J. Proteome Res.* **7**, 1994-2006 (2008).
207. Yan, W. et al. Systematic comparison of the human saliva and plasma proteomes. *Proteomics Clin Appl* **3**, 116-134 (2009).
208. Drake, R.R., Cazare, L.H., Semmes, O.J. & Wadsworth, J.T. Serum, salivary and tissue proteomics for discovery of biomarkers for head and neck cancers. *Expert Rev Mol Diagn* **5**, 93-100 (2005).
209. Amado, F.M., Vitorino, R.M., Domingues, P.M., Lobo, M.J. & Duarte, J.A. Analysis of the human saliva proteome. *Expert Rev Proteomics* **2**, 521-539 (2005).
210. Hu, S., Loo, J.A. & Wong, D.T. Human body fluid proteome analysis. *Proteomics* **6**, 6326-6353 (2006).
211. Xie, H. et al. Proteomics Analysis of Cells in Whole Saliva from Oral Cancer Patients via Value-added Three-dimensional Peptide Fractionation and Tandem Mass Spectrometry. *Mol Cell Proteomics* **7**, 486-498 (2008).
212. Hu, S. et al. Large-scale identification of proteins in human salivary proteome by liquid chromatography/mass spectrometry and two-dimensional gel electrophoresis-mass spectrometry. *Proteomics.* **5**, 1714-1728 (2005).
213. Xie, H., Rhodus, N.L., Griffin, R.J., Carlis, J.V. & Griffin, T.J. A Catalogue of Human Saliva Proteins Identified by Free Flow Electrophoresis-based Peptide Separation and Tandem Mass Spectrometry. *Mol Cell Proteomics* **4**, 1826-1830 (2005).
214. Turnbaugh, P.J. et al. The human microbiome project. *Nature* **449**, 804-810 (2007).
215. HOMD Human Oral Microbiome Database, . *www.homd.org* (2008).
216. van Winkelhoff, A.J., Loos, B.G., van der Reijden, W.A. & van der Velden, U. Porphyromonas gingivalis, Bacteroides forsythus and other putative periodontal pathogens in subjects with and without periodontal destruction. *J Clin Periodontol* **29**, 1023-1028 (2002).
217. Pratten, J., Wilson, M. & Spratt, D.A. Characterization of in vitro oral bacterial biofilms by traditional and molecular methods. *Oral Microbiol Immunol* **18**, 45-49 (2003).
218. Wilson, M. Use of constant depth film fermentor in studies of biofilms of oral bacteria. *Methods Enzymol* **310**, 264-279 (1999).
219. Pratten, J., Smith, A.W. & Wilson, M. Response of single species biofilms and microcosm dental plaques to pulsing with chlorhexidine. *J Antimicrob Chemother* **42**, 453-459 (1998).
220. Zhang, X., Li, L., Wei, D., Yap, Y. & Chen, F. Moving cancer diagnostics from bench to bedside. *Trends in Biotechnology* **25**, 166-173 (2007).
221. Abdullah-Sayani A, B.-d.-M.J., van de Vijver MJ. Technology Insight: tuning into the genetic orchestra using microarrays--limitations of DNA microarrays in clinical practice. *Nat Clin Pract Oncol* **3**, 501-516 (2006).
222. Sandvik, A.K. et al. Gene expression analysis and clinical diagnosis. *Clinica Chimica Acta* **363**, 157-164 (2006).
223. Yang, J. et al. Diagnosis of liver cancer using HPLC-based metabonomics avoiding false-positive result from hepatitis and hepatocirrhosis diseases. *Journal of Chromatography B* **813**, 59-65 (2004).
224. Odunsi, K. et al. Detection of epithelial ovarian cancer using 1H-NMR-based metabonomics.

Int J Cancer **113**, 782-788 (2005).
225. Kobata, A. & Amano, J. Altered glycosylation of proteins produced by malignant cells, and application for the diagnosis and immunotherapy of tumours. *Immunol Cell Biol* **83**, 429-439 (2005).
226. Umezu-Goto, M. et al. Lysophosphatidic acid production and action: validated targets in cancer? *J Cell Biochem* **92**, 1115-1140 (2004).
227. Wenk, M.R. The emerging field of lipidomics. *Nat Rev Drug Discov* **4**, 594-610 (2005).
228. Li, R. et al. Identification of putative oncogenes in lung adenocarcinoma by a comprehensive functional genomic approach. *Oncogene* **25**, 2628-2635 (2006).
229. Burtis, C.A., Ashwood, E. R., Bruns, E. E. Tietz Textbook of Clinical Chemistry. *Elsevier Sauders Co Philadelphia* (2005).
230. Wong, D.T. Salivary diagnostics powered by nanotechnologies, proteomics and genomics. *J Am Dent Assoc* **137**, 313-321 (2006).
231. Liao, P.H., Chang, Y.C., Huang, M.F., Tai, K.W. & Chou, M.Y. Mutation of p53 gene codon 63 in saliva as a molecular marker for oral squamous cell carcinomas. *Oral Oncol* **36**, 272-276 (2000).
232. Warnakulasuriya, S., Soussi, T., Maher, R., Johnson, N. & Tavassoli, M. Expression of p53 in oral squamous cell carcinoma is associated with the presence of IgG and IgA p53 autoantibodies in sera and saliva of the patients. *J Pathol* **192**, 52-57 (2000).
233. Eley, B.C., SW Advances in periodontal diagnosis 6. Proteolytic and hydrolytic enzymes of inflammatory cell origin. *Br Dent J* **184**, 268-271 (1999).
234. Feist, E. & Hansen, A. [Critical overview of outcome parameters for patients with primary Sjogren's syndrome]. *Z Rheumatol* **69**, 25-31 (2010).
235. Rujner, J. et al. Serum and salivary antigliadin antibodies and serum IgA anti-endomysium antibodies as a screening test for coeliac disease. *Acta Paediatr* **85**, 814-817 (1996).
236. Hu, S. et al. Salivary proteomic and genomic biomarkers for primary Sjogren's syndrome. *Arthritis Rheum* **56**, 3588-3600 (2007).
237. Tamashiro, H. & Constantine, N.T. Serological diagnosis of HIV infection using oral fluid samples. *Bull World Health Organ* **72**, 135-143 (1994).
238. Parisi, M.R. et al. Offer of rapid testing and alternative biological samples as practical tools to implement HIV screening programs. *New Microbiol* **32**, 391-396 (2009).
239. Moorthy, M., Daniel, H.D., Kurian, G. & Abraham, P. An evaluation of saliva as an alternative to plasma for the detection of hepatitis C virus antibodies. *Indian J Med Microbiol* **26**, 327-332 (2008).
240. Hodinka, R.L., Nagashunmugam, T. & Malamud, D. Detection of Human Immunodeficiency Virus Antibodies in Oral Fluids. *Clin. Diagn. Lab. Immunol.* **5**, 419-426 (1998).
241. Fisker, N., Georgsen, J., Stolborg, T., Khalil, M.R. & Christensen, P.B. Low hepatitis B prevalence among pre-school children in Denmark: saliva anti-HBc screening in day care centres. *J Med Virol* **68**, 500-504 (2002).
242. Nigatu, W. et al. Measles virus strains circulating in Ethiopia in 1998-1999: molecular characterisation using oral fluid samples and identification of a new genotype. *J Med Virol* **65**, 373-380 (2001).
243. Cone, E.J. Legal, workplace, and treatment drug testing with alternate biological matrices on a global scale. *Forensic Science International* **121**, 7-15 (2001).
244. Chatterton, R.T., Jr., Vogelsong, K.M., Lu, Y.C., Ellman, A.B. & Hudgens, G.A. Salivary alpha-amylase as a measure of endogenous adrenergic activity. *Clin Physiol* **16**, 433-448 (1996).
245. Lu, Y., Bentley, G.R., Gann, P.H., Hodges, K.R. & Chatterton, R.T. Salivary estradiol and progesterone levels in conception and nonconception cycles in women: evaluation of a new assay for salivary estradiol. *Fertil Steril* **71**, 863-868 (1999).
246. Navarro, M.A. et al. Epidermal growth factor in plasma and saliva of patients with active breast cancer and breast cancer patients in follow-up compared with healthy women. *Breast Cancer Res Treat* **42**, 83-86 (1997).
247. Chen, D.X., Schwartz, P.E. & Li, F.Q. Saliva and serum CA 125 assays for detecting malignant ovarian tumors. *Obstet Gynecol* **75**, 701-704 (1990).
248. Streckfus, C., Bigler, L. & O'Bryan, T. Aging and salivary cytokine concentrations as predictors of whole saliva flow rates among women: a preliminary study. *Gerontology* **48**, 282-288 (2002).

249. Lawrence, H.P. Salivary markers of systemic disease: noninvasive diagnosis of disease and monitoring of general health. *J Can Dent Assoc* **68**, 170-174 (2002).
250. Forde, M.D., Koka, S., Eckert, S.E., Carr, A.B. & Wong, D.T. Systemic assessments utilizing saliva: part 1 general considerations and current assessments. *Int J Prosthodont* **19**, 43-52 (2006).
251. Koka, S., Forde, M.D. & Khosla, S. Systemic assessments utilizing saliva: part 2 osteoporosis and use of saliva to measure bone turnover. *Int J Prosthodont* **19**, 53-60 (2006).
252. Mandel, I.D. Salivary diagnosis: promises, promises. *Ann N Y Acad Sci* **694**, 1-10 (1993).
253. Mandel, I.D. A contemporary view of salivary research. *Crit Rev Oral Biol Med* **4**, 599-604 (1993).
254. Slavkin, H.C. Toward molecularly based diagnostics for the oral cavity. *J Am Dent Assoc* **129**, 1138-1143 (1998).
255. Crouch, D.J., Walsh, J.M., Cangianelli, L. & Quintela, O. Laboratory evaluation and field application of roadside oral fluid collectors and drug testing devices. *Ther Drug Monit* **30**, 188-195 (2008).
256. Walsh, J.M. New technology and new initiatives in U.S. workplace testing. *Forensic Sci Int* **174**, 120-124 (2008).
257. Gao, K. et al. Systemic disease-induced salivary biomarker profiles in mouse models of melanoma and non-small cell lung cancer. *PLoS One* **4**, e5875 (2009).
258. Scheper, H.J. & Brand, H.S. Oral aspects of Crohn's disease. *Int Dent J* **52**, 163-172 (2002).
259. Newsom-Davis, J. Lambert-Eaton myasthenic syndrome. *Rev Neurol (Paris)* **160**, 177-180 (2004).
260. Rayman, S., Dincer, E. & Almas, K. Xerostomia. Diagnosis and management in dental practice. *N Y State Dent J* **76**, 24-27 (2010).
261. Knas, M. et al. Saliva of patients with Type 1 diabetes: effect of smoking on activity of lysosomal exoglycosidases. *Oral Diseases* **12**, 278-282 (2006).
262. Fields, R. The measurement of amino groups in proteins and peptides. *Biochem. J.* **124**, 581-590 (1971).
263. Wessel, D. & Flugge, U.I. A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids. *Anal. Biochem.* **138**, 141-143 (1984).
264. Rappsilber, J., Mann, M. & Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2**, 1896-1906 (2007).
265. Shevchenko, A., Wilm, M., Vorm, O. & Mann, M. Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. *Anal. Chem.* **68**, 850-858 (1996).
266. Haas, W. et al. Optimization and use of peptide mass measurement accuracy in shotgun proteomics. *Mol. Cell. Proteomics.* **5**, 1326-1337 (2006).
267. Kersey, P.J. et al. The International Protein Index: an integrated database for proteomics experiments. *Proteomics.* **4**, 1985-1988 (2004).
268. Elias, J.E. & Gygi, S.P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods.* **4**, 207-214 (2007).
269. Bakalarski, C.E. et al. The impact of Peptide abundance and dynamic range on stable-isotope-based quantitative proteomic analyses. *J. Proteome. Res.* **7**, 4756-4765 (2008).
270. Landy, A. Dynamic, structural, and regulatory aspects of lambda site-specific recombination. *Annu Rev Biochem* **58**, 913-949 (1989).
271. Hartley, J.L., Temple, G.F. & Brasch, M.A. DNA cloning using in vitro site-specific recombination. *Genome Res* **10**, 1788-1795 (2000).
272. Invitrogen *Invitrogen Life Technologies, Instruction manual for Gateway pENTR Vectors* (2004).
273. Qiagen *QIAprep Miniprep Handbook* (2006).
274. Invitrogen *Lipofectamine 2000 protocol* (2006).
275. Schagger, H. & von Jagow, G. Blue native electrophoresis for isolation of membrane protein complexes in enzymatically active form. *Anal Biochem* **199**, 223-231 (1991).
276. Schagger, H. Blue-native gels to isolate protein complexes from mitochondria. *Methods Cell Biol* **65**, 231-244 (2001).
277. Kloetzel, P.M. Antigen processing by the proteasome. *Nat Rev Mol Cell Biol* **2**, 179-187 (2001).
278. Roelofs, J. et al. Chaperone-mediated pathway of proteasome regulatory particle assembly.

Nature **459**, 861-865 (2009).
279. Li, S.J. et al. Sys-BodyFluid: a systematical database for human body fluid proteome research. *Nucleic Acids Res* **37**, D907-912 (2009).
280. Zhang, Y. et al. MAPU: Max-Planck Unified database of organellar, cellular, tissue and body fluid proteomes. *Nucleic Acids Res* **35**, D771-779 (2007).
281. Binz, P.A., Hochstrasser, D.F. & Appel, R.D. Mass spectrometry-based proteomics: current status and potential use in clinical chemistry. *Clin Chem Lab Med* **41**, 1540-1551 (2003).
282. Fusaro, V.A. & Stone, J.H. Mass spectrometry-based proteomics and analyses of serum: a primer for the clinical investigator. *Clin Exp Rheumatol* **21**, S3-14 (2003).
283. Defabianis, P. & Re, F. [The role of saliva in maintaining oral health]. *Minerva Stomatol* **52**, 301-308 (2003).
284. Turner, R.J. & Sugiya, H. Understanding salivary fluid and protein secretion. *Oral Dis* **8**, 3-11 (2002).
285. Messana, I., Inzitari, R., Fanali, C., Cabras, T. & Castagnola, M. Facts and artifacts in proteomics of body fluids. What proteomics of saliva is telling us? *J Sep Sci* **31**, 1948-1963 (2008).
286. Dalwai, F., Spratt, D.A. & Pratten, J. Modeling Shifts in Microbial Populations Associated with Health or Disease. *Appl. Environ. Microbiol.* **72**, 3678-3684 (2006).
287. Darveau, R.P., Tanner, A. & Page, R.C. The microbial challenge in periodontitis. *Periodontol 2000* **14**, 12-32 (1997).
288. Scannapieco, F.A. Saliva-bacterium interactions in oral microbial ecology. *Crit Rev Oral Biol Med* **5**, 203-248 (1994).
289. Rudney, J.D. Does variability in salivary protein concentrations influence oral microbial ecology and oral health? *Crit Rev Oral Biol Med* **6**, 343-367 (1995).
290. Drzymala, L., Castle, A., Cheung, J.C. & Bennick, A. Cellular phosphorylation of an acidic proline-rich protein, PRP1, a secreted salivary phosphoprotein. *Biochemistry* **39**, 2023-2031 (2000).
291. Salih, E. *Grant 1R21DE018448-01 A2 from National Institute of Dental and Craniofacial Research*.
292. Ranney, R.R. Diagnosis of periodontal diseases. *Adv Dent Res* **5**, 21-36 (1991).
293. Tenovuo, J., Jentsch, H., Soukka, T. & Karhuvaara, L. Antimicrobial factors of saliva in relation to dental caries and salivary levels of mutans streptococci. *J Biol Buccale* **20**, 85-90 (1992).
294. Biesbrock, A.R., Reddy, M.S. & Levine, M.J. Interaction of a salivary mucin-secretory immunoglobulin A complex with mucosal pathogens. *Infect Immun* **59**, 3492-3497 (1991).
295. Smith, C.A., Want, E.J., O'Maille, G., Abagyan, R. & Siuzdak, G. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Anal. Chem.* **78**, 779-787 (2006).
296. Ruijken, M. MsMetrix. *www.msmetrix.com* (2006).
297. (Progenesis nonlinear Dynamics www.nonlinear.com 2008).
298. Dalwai, F., Spratt, D.A. & Pratten, J. Use of Quantitative PCR and Culture Methods To Characterize Ecological Flux in Bacterial Biofilms. *J. Clin. Microbiol.* **45**, 3072-3076 (2007).
299. Nikolsky, Y., Kirillov, E., Zuev, R., Rakhmatulin, E. & Nikolskaya, T. Functional analysis of OMICs data and small molecule compounds in an integrated "knowledge-based" platform. *Methods Mol Biol* **563**, 177-196 (2009).
300. Johnson, R.J. et al. Analysis of gene ontology features in microarray data using the Proteome BioKnowledge Library. *In Silico Biol* **5**, 389-399 (2005).
301. Chow, S.R., P. Constructing Area-Proportional Venn and Euler Diagrmas with Three Circles. *Euler Diagrams Workshop. Paris* (2005).
302. Rowat, J.S. & Squier, C.A. Rates of epithelial cell proliferation in the oral mucosa and skin of the tamarin monkey (Saguinus fuscicollis). *J Dent Res* **65**, 1326-1331 (1986).
303. Villen, J. & Gygi, S.P. The SCX/IMAC enrichment approach for global phosphorylation analysis by mass spectrometry. *Nat. Protoc.* **3**, 1630-1638 (2008).
304. Cavicchioli, R., Curmi, P.M., Saunders, N. & Thomas, T. Pathogenic archaea: do they exist? *Bioessays* **25**, 1119-1128 (2003).
305. Farhoud, M.H. et al. Protein complexes in the archaeon Methanothermobacter thermautotrophicus analyzed by blue native/SDS-PAGE and mass spectrometry. *Mol Cell Proteomics* **4**, 1653-1663 (2005).

306. Wessels, H.J. et al. LC-MS/MS as an alternative for SDS-PAGE in blue native analysis of protein complexes. *Proteomics* **9**, 4221-4228 (2009).
307. Husi, H. Proteomics Analysis Datbase PADB. *www.PADB.org* (2008).
308. Beausoleil, S.A. et al. Large-scale characterization of HeLa cell nuclear phosphoproteins. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 12130-12135 (2004).
309. Villen, J., Beausoleil, S.A., Gerber, S.A. & Gygi, S.P. Large-scale phosphorylation analysis of mouse liver. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 1488-1493 (2007).
310. Huttlin, E. A Tissue-Specific Atlas of Mouse Protein Phosphorylation and Expression. *Submitted.* (2010).
311. Amanchy, R. et al. A curated compendium of phosphorylation motifs. *Nat Biotechnol* **25**, 285-286 (2007).
312. Kemp, B.E., Parker, M.W., Hu, S., Tiganis, T. & House, C. Substrate and pseudosubstrate interactions with protein kinases: determinants of specificity. *Trends Biochem Sci* **19**, 440-444 (1994).
313. Lorenz, M., Saretzki, G., Sitte, N., Metzkow, S. & von Zglinicki, T. BJ fibroblasts display high antioxidant capacity and slow telomere shortening independent of hTERT transfection. *Free Radic Biol Med* **31**, 824-831 (2001).
314. Takai, H., Smogorzewska, A. & de Lange, T. DNA damage foci at dysfunctional telomeres. *Curr Biol* **13**, 1549-1556 (2003).
315. d'Adda di Fagagna, F. et al. A DNA damage checkpoint response in telomere-initiated senescence. *Nature* **426**, 194-198 (2003).
316. Martin-Ruiz, C. et al. Stochastic variation in telomere shortening rate causes heterogeneity of human fibroblast replicative life span. *J Biol Chem* **279**, 17826-17833 (2004).
317. Manning, G., Plowman, G.D., Hunter, T. & Sudarsanam, S. Evolution of protein kinase signaling from yeast to man. *Trends Biochem Sci* **27**, 514-520 (2002).
318. Gallez, F., Fadel, M., Scruel, O., Cantraine, F. & Courtois, P. Salivary biomass assessed by bioluminescence ATP assay related to (bacterial and somatic) cell counts. *Cell Biochem Funct* **18**, 103-108 (2000).
319. Kannan, N., Taylor, S.S., Zhai, Y., Venter, J.C. & Manning, G. Structural and functional diversity of the microbial kinome. *PLoS Biol* **5**, e17 (2007).
320. Su, A.I. et al. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* **101**, 6062-6067 (2004).
321. Bantscheff, M. et al. Quantitative chemical proteomics reveals mechanisms of action of clinical ABL kinase inhibitors. *Nat. Biotechnol.* **25**, 1035-1044 (2007).
322. Rennefahrt, U.E. et al. Specificity profiling of Pak kinases allows identification of novel phosphorylation sites. *J Biol Chem* **282**, 15667-15678 (2007).
323. Zhu, G. et al. A single pair of acidic residues in the kinase major groove mediates strong substrate preference for P-2 or P-5 arginine in the AGC, CAMK, and STE kinase families. *J Biol Chem* **280**, 36372-36379 (2005).
324. Bagrodia, S. & Cerione, R.A. Pak to the future. *Trends Cell Biol* **9**, 350-355 (1999).
325. Allen, K.M. et al. PAK3 mutation in nonsyndromic X-linked mental retardation. *Nat Genet* **20**, 25-30 (1998).
326. Nigg, E.A. Polo-like kinases: positive regulators of cell division from start to finish. *Curr Opin Cell Biol* **10**, 776-783 (1998).
327. Hauf, S., Waizenegger, I.C. & Peters, J.M. Cohesin cleavage by separase required for anaphase and cytokinesis in human cells. *Science* **293**, 1320-1323 (2001).
328. Nakajima, H., Toyoshima-Morimoto, F., Taniguchi, E. & Nishida, E. Identification of a consensus motif for Plk (Polo-like kinase) phosphorylation reveals Myt1 as a Plk1 substrate. *J Biol Chem* **278**, 25277-25280 (2003).
329. Yarm, F.R. Plk phosphorylation regulates the microtubule-stabilizing protein TCTP. *Mol Cell Biol* **22**, 6209-6221 (2002).
330. Anderson, M. et al. Plo1(+) regulates gene transcription at the M-G(1) interval during the fission yeast mitotic cell cycle. *EMBO J* **21**, 5745-5755 (2002).
331. Cho, R.J. et al. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell* **2**, 65-73 (1998).
332. Hansen, D.V., Loktev, A.V., Ban, K.H. & Jackson, P.K. Plk1 regulates activation of the anaphase promoting complex by phosphorylating and triggering SCFbetaTrCP-dependent

destruction of the APC Inhibitor Emi1. *Mol Biol Cell* **15**, 5623-5634 (2004).
333. Watanabe, N. et al. M-phase kinases induce phospho-dependent ubiquitination of somatic Wee1 by SCFbeta-TrCP. *Proc Natl Acad Sci U S A* **101**, 4419-4424 (2004).
334. Zhou, T., Aumais, J.P., Liu, X., Yu-Lee, L.Y. & Erikson, R.L. A role for Plk1 phosphorylation of NudC in cytokinesis. *Dev Cell* **5**, 127-138 (2003).
335. Neef, R. et al. Phosphorylation of mitotic kinesin-like protein 2 by polo-like kinase 1 is required for cytokinesis. *J Cell Biol* **162**, 863-875 (2003).
336. Casenghi, M. et al. Polo-like kinase 1 regulates Nlp, a centrosome protein involved in microtubule nucleation. *Dev Cell* **5**, 113-125 (2003).
337. Toyoshima-Morimoto, F., Taniguchi, E. & Nishida, E. Plk1 promotes nuclear translocation of human Cdc25C during prophase. *EMBO Rep* **3**, 341-348 (2002).
338. Laoukili, J. et al. FoxM1 is required for execution of the mitotic programme and chromosome stability. *Nat Cell Biol* **7**, 126-136 (2005).
339. Major, M.L., Lepe, R. & Costa, R.H. Forkhead box M1B transcriptional activity requires binding of Cdk-cyclin complexes for phosphorylation-dependent recruitment of p300/CBP coactivators. *Mol Cell Biol* **24**, 2649-2661 (2004).
340. Wang, I.C. et al. Forkhead box M1 regulates the transcriptional network of genes essential for mitotic progression and genes encoding the SCF (Skp2-Cks1) ubiquitin ligase. *Mol Cell Biol* **25**, 10875-10894 (2005).
341. Fu, Z. et al. Plk1-dependent phosphorylation of FoxM1 regulates a transcriptional programme required for mitotic progression. *Nat Cell Biol* **10**, 1076-1082 (2008).
342. Lowery, D.M. et al. Proteomic screen defines the Polo-box domain interactome and identifies Rock2 as a Plk1 substrate. *EMBO J* **26**, 2262-2273 (2007).
343. Elia, A.E., Cantley, L.C. & Yaffe, M.B. Proteomic screen finds pSer/pThr-binding domain localizing Plk1 to mitotic substrates. *Science.* **299**, 1228-1231 (2003).
344. Gumireddy, K. et al. ON01910, a non-ATP-competitive small molecule inhibitor of Plk1, is a potent anticancer agent. *Cancer Cell* **7**, 275-286 (2005).
345. Gaudet, S., Branton, D. & Lue, R.A. Characterization of PDZ-binding kinase, a mitotic kinase. *Proc Natl Acad Sci U S A* **97**, 5167-5172 (2000).
346. Zykova, T.A. et al. Lymphokine-activated killer T-cell-originated protein kinase phosphorylation of histone H2AX prevents arsenite-induced apoptosis in RPMI7951 melanoma cells. *Clin Cancer Res* **12**, 6884-6893 (2006).
347. Ayllon, V. & O'Connor, R. PBK/TOPK promotes tumour cell proliferation through p38 MAPK activity and regulation of the DNA damage response. *Oncogene* **26**, 3451-3461 (2007).
348. Nandi, A.K., Ford, T., Fleksher, D., Neuman, B. & Rapoport, A.P. Attenuation of DNA damage checkpoint by PBK, a novel mitotic kinase, involves protein-protein interaction with tumor suppressor p53. *Biochem Biophys Res Commun* **358**, 181-188 (2007).
349. Bishop, A.C. et al. A chemical switch for inhibitor-sensitive alleles of any protein kinase. *Nature* **407**, 395-401 (2000).
350. Bostrom, E.A. et al. Salivary Resistin Reflects Local Inflammation in Sjogren's Syndrome. *J Rheumatol* (2008).
351. vS Agarwal, N.P., LP Johns Differential expression of IL-1β, TNF-, IL-6, and IL-8 in human monocytes in response to lipopolysaccharides from different microbes. *J Dent Res* **74**, 1057-1065 (1995).
352. Eley, B.M. & Cox, S.W. Advances in periodontal diagnosis. 1. Traditional clinical methods of diagnosis. *Br Dent J* **184**, 12-16 (1998).
353. Socransky SS, H.A. The bacterial etiology of destructive periodontal disease: current concepts. *J Periodontol* **63**, 322-337 (1992).
354. Dayan, S., Stashenko, P., Niederman, R. & Kupper, T.S. Oral Epithelial Overexpression of IL-1{alpha} Causes Periodontal Disease. *J Dent Res* **83**, 786-790 (2004).
355. Schroeder HE, M.-P.S., Page R Correlated morphometric and biochemical analysis of gingival tissue in early chronic gingivitis in man. *Arch Oral Biol* **18**, 899-834 (1973).
356. Rai, B., Kharb, S., Jain, R. & Anand, S.C. Biomarkers of periodontitis in oral fluids. *J Oral Sci* **50**, 53-56 (2008).
357. Todorovic, T. et al. Salivary enzymes and periodontal disease. *Med Oral Patol Oral Cir Bucal. 2006 Mar 1;11(2):E115-9.*
358. Miller, C.S., King, C.P., Jr., Langub, M.C., Kryscio, R.J. & Thomas, M.V. Salivary

biomarkers of existing periodontal disease: A cross-sectional study. *J Am Dent Assoc* **137**, 322-329 (2006).
359. Wu, Y., Shu, R., Luo, L.J., Ge, L.H. & Xie, Y.F. Initial comparison of proteomic profiles of whole unstimulated saliva obtained from generalized aggressive periodontitis patients and healthy control subjects. *J Periodontal Res* **44**, 636-644 (2009).
360. Goncalves Lda, R. et al. Comparative proteomic analysis of whole saliva from chronic periodontitis patients. *J Proteomics* **73**, 1334-1341 (2010).
361. Haigh, B.J. et al. Alterations in the salivary proteome associated with periodontitis. *J Clin Periodontol* **37**, 241-247 (2010).
362. Bostanci, N. et al. Application of label-free absolute quantitative proteomics in human gingival crevicular fluid by LC/MS E (gingival exudatome). *J Proteome Res* **9**, 2191-2199 (2010).
363. Dea, M.K. et al. Albumin binding of acylated insulin (NN304) does not deter action to stimulate glucose uptake. *Diabetes* **51**, 762-769 (2002).
364. Offenbacher, S. et al. Gingival transcriptome patterns during induction and resolution of experimental gingivitis in humans. *J Periodontol* **80**, 1963-1982 (2009).
365. Barros, S.P. et al. Triclosan inhibition of acute and chronic inflammatory gene pathways. *J Clin Periodontol* **37**, 412-418 (2010).
366. Zappacosta, B. et al. Salivary thiols and enzyme markers of cell damage in periodontal disease. *Clin Biochem* **40**, 661-665 (2007).
367. Vardar-Sengul, S., Arora, S., Baylas, H. & Mercola, D. Expression profile of human gingival fibroblasts induced by interleukin-1beta reveals central role of nuclear factor-kappa B in stabilizing human gingival fibroblasts during inflammation. *J Periodontol* **80**, 833-849 (2009).
368. Chen, H.Y., Cox, S.W. & Eley, B.M. Cathepsin B, alpha2-macroglobulin and cystatin levels in gingival crevicular fluid from chronic periodontitis patients. *J Clin Periodontol* **25**, 34-41 (1998).
369. Dickinson, D.P. Cysteine peptidases of mammals: their biological roles and potential effects in the oral cavity and other tissues in health and disease. *Crit Rev Oral Biol Med* **13**, 238-275 (2002).
370. Tervahartiala, T. et al. Cathepsin G in gingival tissue and crevicular fluid in adult periodontitis. *J Clin Periodontol* **23**, 68-75 (1996).
371. Sorsa, T. et al. Scientific basis of a matrix metalloproteinase-8 specific chair-side test for monitoring periodontal and peri-implant health and disease. *Ann N Y Acad Sci* **878**, 130-140 (1999).
372. Yamalik, N., Caglayan, F., Kilinc, K., Kilinc, A. & Tumer, C. The importance of data presentation regarding gingival crevicular fluid myeloperoxidase and elastase-like activity in periodontal disease and health status. *J Periodontol* **71**, 460-467 (2000).
373. Perretti, M. & D'Acquisto, F. Annexin A1 and glucocorticoids as effectors of the resolution of inflammation. *Nat Rev Immunol* **9**, 62-70 (2009).
374. Parente, L. & Solito, E. Annexin 1: more than an anti-phospholipase protein. *Inflamm Res* **53**, 125-132 (2004).
375. Marenholz, I., Heizmann, C.W. & Fritz, G. S100 proteins in mouse and man: from evolution to function and pathology (including an update of the nomenclature). *Biochem Biophys Res Commun* **322**, 1111-1122 (2004).
376. Ebersole, J.L., Machen, R.L., Steffen, M.J. & Willmann, D.E. Systemic acute-phase reactants, C-reactive protein and haptoglobin, in adult periodontitis. *Clin Exp Immunol* **107**, 347-352 (1997).
377. Lundy, F.T., Orr, D.F., Shaw, C., Lamey, P.J. & Linden, G.J. Detection of individual human neutrophil alpha-defensins (human neutrophil peptides 1, 2 and 3) in unfractionated gingival crevicular fluid--a MALDI-MS approach. *Mol Immunol* **42**, 575-579 (2005).
378. Brogden, K.A. et al. Defensin-induced adaptive immunity in mice and its potential in preventing periodontal disease. *Oral Microbiol Immunol* **18**, 95-99 (2003).
379. Udby, L., Calafat, J., Sorensen, O.E., Borregaard, N. & Kjeldsen, L. Identification of human cysteine-rich secretory protein 3 (CRISP-3) as a matrix protein in a subset of peroxidase-negative granules of neutrophils and in the granules of eosinophils. *J Leukoc Biol* **72**, 462-469 (2002).
380. Laine, M. et al. Low salivary dehydroepiandrosterone and androgen-regulated cysteine-rich

secretory protein 3 levels in Sjogren's syndrome. *Arthritis Rheum* **56**, 2575-2584 (2007).
381. Ye, H. et al. Transcriptomic dissection of tongue squamous cell carcinoma. *BMC Genomics* **9**, 69 (2008).
382. Kosari, F., Asmann, Y.W., Cheville, J.C. & Vasmatzis, G. Cysteine-rich secretory protein-3: a potential biomarker for prostate cancer. *Cancer Epidemiol Biomarkers Prev* **11**, 1419-1426 (2002).
383. Liao, Q. et al. Preferential expression of cystein-rich secretory protein-3 (CRISP-3) in chronic pancreatitis. *Histol Histopathol* **18**, 425-433 (2003).
384. Honore, B. The rapidly expanding CREC protein family: members, localization, function, and role in disease. *Bioessays* **31**, 262-277 (2009).
385. Cho, J.H. et al. Pleiotropic roles of calumenin (calu-1), a calcium-binding ER luminal protein, in Caenorhabditis elegans. *FEBS Lett* **583**, 3050-3056 (2009).
386. Enyedi, B., Varnai, P. & Geiszt, M. Redox State of the Endoplasmic Reticulum Is Controlled by Ero1L-alpha and Intraluminal Calcium. *Antioxid Redox Signal* (2010).
387. Li, G. et al. Role of ERO1-alpha-mediated stimulation of inositol 1,4,5-triphosphate receptor activity in endoplasmic reticulum stress-induced apoptosis. *J Cell Biol* **186**, 783-792 (2009).
388. Sheng, W.Y. & Wang, T.C. Proteomic analysis of the differential protein expression reveals nuclear GAPDH in activated T lymphocytes. *PLoS One* **4**, e6322 (2009).
389. Ikeda, K. et al. Glia maturation factor-gamma is preferentially expressed in microvascular endothelial and inflammatory cells and modulates actin cytoskeleton reorganization. *Circ Res* **99**, 424-433 (2006).
390. Guven, O. & De Visscher, J.G. Salivary IgA in periodontal disease. *J Periodontol* **53**, 334-335 (1982).
391. Shapiro, L., Goldman, H. & Bloom, A. Sulcular exudate flow in gingival inflammation. *J Periodontol* **50**, 301-304 (1979).
392. Fourie, A.M. Modulation of inflammatory disease by inhibitors of leukotriene A4 hydrolase. *Curr Opin Investig Drugs* **10**, 1173-1182 (2009).
393. Ramseier, C.A. et al. Identification of pathogen and host-response markers correlated with periodontal disease. *J Periodontol* **80**, 436-446 (2009).
394. Shelburne, C.E., Prabhu, A., Gleason, R.M., Mullally, B.H. & Coulter, W.A. Quantitation of Bacteroides forsythus in subgingival plaque comparison of immunoassay and quantitative polymerase chain reaction. *J Microbiol Methods* **39**, 97-107 (2000).
395. Mullally, B.H., Dace, B., Shelburne, C.E., Wolff, L.F. & Coulter, W.A. Prevalence of periodontal pathogens in localized and generalized forms of early-onset periodontitis. *J Periodontal Res* **35**, 232-241 (2000).
396. Persson, G.R. et al. Tannerella forsythia and Pseudomonas aeruginosa in subgingival bacterial samples from parous women. *J Periodontol* **79**, 508-516 (2008).
397. Haffajee, A.D. & Socransky, S.S. Microbiology of periodontal diseases: introduction. *Periodontol 2000* **38**, 9-12 (2005).
398. Holt, S.C. & Ebersole, J.L. Porphyromonas gingivalis, Treponema denticola, and Tannerella forsythia: the "red complex", a prototype polybacterial pathogenic consortium in periodontitis. *Periodontol 2000* **38**, 72-122 (2005).
399. Aas, J.A. et al. Subgingival plaque microbiota in HIV positive patients. *J Clin Periodontol* **34**, 189-195 (2007).
400. Daines, D.A. et al. Haemophilus influenzae luxS mutants form a biofilm and have increased virulence. *Microb Pathog* **39**, 87-96 (2005).
401. Yarwood, J.M. & Schlievert, P.M. Quorum sensing in Staphylococcus infections. *J Clin Invest* **112**, 1620-1625 (2003).
402. Okuda, K., Kimizuka, R., Abe, S., Kato, T. & Ishihara, K. Involvement of periodontopathic anaerobes in aspiration pneumonia. *J Periodontol* **76**, 2154-2160 (2005).
403. Tada, A., Watanabe, T., Yokoe, H., Hanada, N. & Tanzawa, H. Oral bacteria influenced by the functional status of the elderly people and the type and quality of facilities for the bedridden. *J Appl Microbiol* **93**, 487-491 (2002).
404. Socransky, S.S., Haffajee, A.D., Cugini, M.A., Smith, C. & Kent, R.L., Jr. Microbial complexes in subgingival plaque. *J Clin Periodontol* **25**, 134-144 (1998).
405. Saba, J.A., McComb, M.E., Potts, D.L., Costello, C.E. & Amar, S. Proteomic mapping of stimulus-specific signaling pathways involved in THP-1 cells exposed to Porphyromonas gingivalis or its purified components. *J Proteome Res* **6**, 2211-2221 (2007).

406. Craig, R.G. et al. Relationship of destructive periodontal disease to the acute-phase response. *J Periodontol* **74**, 1007-1016 (2003).
407. Nakagawa, I. et al. Distribution and molecular characterization of Porphyromonas gingivalis carrying a new type of fimA gene. *J Clin Microbiol* **38**, 1909-1914 (2000).
408. Kitahara, M., Sakamoto, M., Ike, M., Sakata, S. & Benno, Y. Bacteroides plebeius sp. nov. and Bacteroides coprocola sp. nov., isolated from human faeces. *Int J Syst Evol Microbiol* **55**, 2143-2147 (2005).
409. Wery, N., Monteil, C., Pourcher, A.M. & Godon, J.J. Human-specific fecal bacteria in wastewater treatment plant effluents. *Water Res* **44**, 1873-1883 (2010).
410. Riep, B. et al. Are putative periodontal pathogens reliable diagnostic markers? *J Clin Microbiol* **47**, 1705-1711 (2009).
411. Colombo, A.P. et al. Subgingival microbiota of Brazilian subjects with untreated chronic periodontitis. *J Periodontol* **73**, 360-369 (2002).
412. Aramaki, M., Nagasawa, T., Koseki, T. & Ishikawa, I. Presence of activated B-1 cells in chronic inflamed gingival tissue. *J Clin Immunol* **18**, 421-429 (1998).
413. Geibel, M.A., Schu, B., Callaway, A.S., Gleissner, C. & Willershausen, B. Polymerase chain reaction-based simultaneous detection of selected bacterial species associated with closed periapical lesions. *Eur J Med Res* **10**, 333-338 (2005).
414. Neil, R.B. & Apicella, M.A. Clinical and laboratory evidence for Neisseria meningitidis biofilms. *Future Microbiol* **4**, 555-563 (2009).
415. Urnowey, S. et al. Temporal activation of anti- and pro-apoptotic factors in human gingival fibroblasts infected with the periodontal pathogen, Porphyromonas gingivalis: potential role of bacterial proteases in host signalling. *BMC Microbiol* **6**, 26 (2006).
416. Stathopoulou, P.G., Benakanakere, M.R., Galicia, J.C. & Kinane, D.F. The host cytokine response to Porphyromonas gingivalis is modified by gingipains. *Oral Microbiol Immunol* **24**, 11-17 (2009).
417. Stathopoulou, P.G. et al. Porphyromonas gingivalis induce apoptosis in human gingival epithelial cells through a gingipain-dependent mechanism. *BMC Microbiol* **9**, 107 (2009).
418. Festa, R.A. et al. Prokayrotic ubiquitin-like protein (Pup) proteome of Mycobacterium tuberculosis. *PLoS One* **5**, e8589 (2010).

Appendices

Proteomics and protein activity profiling:

An investigation into the salivary proteome and kinase activities in various systems using mass spectrometry

Fiona E. McAllister

The University of Edinburgh

Doctor of Philosophy by Research

June 2010

Table of Contents for Appendices

A. Collaborative work

Aside from the two main projects on saliva and KAYAK, I was involved with various collaborative projects. A short description of my contribution to the main collaborative projects is given below.

A.1 Chaperone-mediated pathway of proteasome regulatory particle assembly

Letter

Nature **459**, 861-865 (11 June 2009) | doi:10.1038/nature08063; Received 24 October 2008; Accepted 15 April 2009; Published online 1 May 2009

Chaperone-mediated pathway of proteasome regulatory particle assembly

Jeroen Roelofs[1], Soyeon Park[1], Wilhelm Haas[1], Geng Tian[1], Fiona E. McAllister[1], Ying Huo[1], Byung-Hoon Lee[1], Fan Zhang[2], Yigong Shi[2], Steven P. Gygi[1] & Daniel Finley[1]

1. Department of Cell Biology, Harvard Medical School, 240 Longwood Avenue, Boston, Massachusetts 02115, USA
2. Department of Molecular Biology, Lewis Thomas Laboratory, Princeton University, Princeton, New Jersey 08544, USA

Correspondence to: Daniel Finley[1] Correspondence and requests for materials should be addressed to D.F. (Email: daniel_finley@hms.harvard.edu).

The proteasome is a protease that controls diverse processes in eukaryotic cells. Its regulatory particle (RP) initiates the degradation of ubiquitin–protein conjugates by unfolding the substrate and translocating it into the proteasome core particle (CP) to be degraded[1]. The RP has 19 subunits, and their pathway of assembly is not understood. Here we show that in the yeast *Saccharomyces cerevisiae* three proteins are found associated with RP but not with the RP–CP holoenzyme: Nas6, Rpn14 and Hsm3. Mutations in the corresponding genes confer proteasome loss-of-function phenotypes, despite their virtual absence from the holoenzyme. These effects result from deficient RP assembly. Thus, Nas6, Rpn14 and Hsm3 are RP chaperones. The RP contains six ATPases–the Rpt proteins–and each RP chaperone binds to the carboxy-terminal domain of a specific Rpt. We show in an accompanying study[2] that RP assembly is templated through the Rpt C termini, apparently by their insertion into binding pockets in the CP. Thus, RP chaperones may regulate proteasome assembly by directly restricting the accessibility of Rpt C termini to the CP. In addition, competition between the RP chaperones and the CP for Rpt engagement may explain the release of RP chaperones as proteasomes mature.

This was a collaboration with Dr Jeroen Roelofs at Harvard Medical School investigating the proteasome. As part of the work on the blue native PAGE gels for the in-gel KAYAK correlation profiling project (Section 2.7.7.1), one of the preliminary steps was to determine what complexes were observed. By plotting the spectral count of each of the proteins against each of the gel fractions it was

possible to detect various protein complexes. All proteasome subunits could be identified in fractions 1 and 2, indicating that completely assembled proteasomes were present in these fractions (Figure A.1). Many of the proteasome subunits were also present in different fractions of lower molecular weight. Further analysis showed that proteasomal proteins known to belong to a specific subcomplex have similar profiles (Figure A.1 a-d). For example, all subunits of the proteasome core particle (CP) were abundant in fractions 1 and 2 as components of assembled proteasomes (26S), but also in fraction 3. Fraction 3 lacks proteasome subunits from the other proteasome subcomplexes, the base and the lid, suggesting that this complex is assembled CP. All lid components showed a similar pattern to one another and could be found in fractions 1, 2 and 4, indicating that in fraction four we identified assembled lid. A base complex was not detected but, interestingly, four proteins that are thought to comprise the BP1 complex, an intermediate in base assembly, were detected in fraction 4 and had similar profiles.

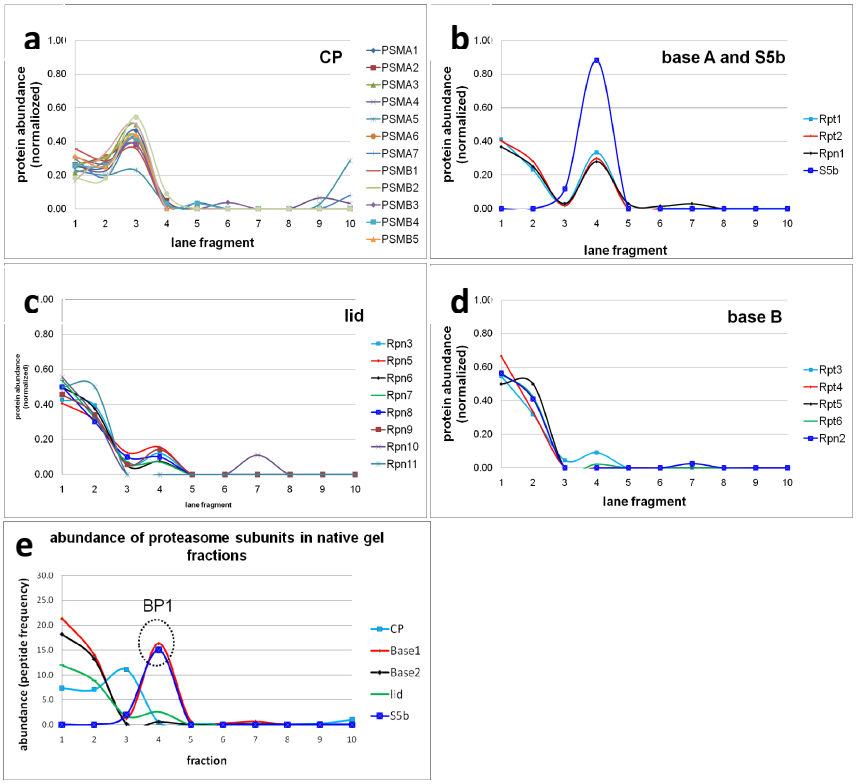


Figure A.1 HeLa cell lysate was run on Blue native gels (Invitrogen). After electrophoresis, gels were stained with Coommassie blue and the lane was divided into ten equal size pieces; all pieces were analyzed by mass spectrometry and protein abundance was determined using spectral counting. Graphs (a-d) display the results for most proteasome subunits and graph (e) comprises an average of all the proteins in the different proteasomal components. Profiles of abundance can be divided into four groups: CP (alpha 1-7 and beta1-7); lid (Rpn3-11); base2 (consisting of Rpt3,4,5,6 and rpn2); Base1 (consisting of Rpt1,2, and Rpn1). All groups are present in fractions 1 and 2, where 26S proteasomes are expected to migrate. Fraction 3 contains a high abundance of CP, while the other groups have a very low abundance, indicating the location of free CP. In fraction 4 there is a specific high abundance of the subunits from group base1, while other subunits (base 2 group) are very low in abundance. S5b (PSMD5) shows a unique profile, being absent for the fractions where 26S proteasomes migrate and only abundant in the fraction where base 1 is present. These data suggest that the BP1 complex identified in yeast (*Saccharomyces cerevisiae*) is conserved in human and that S5b is indeed the ortholog of Hsm3.

Subsequent to publication of the paper, further work was carried out to investigate the BP1 complex in more detail. 26S proteasome consumes BP1 and therefore it would be expected that the turnover of BP1 would be faster than that of the 26S proteasomes. It has previously been shown by pulse-labelling in yeast that ^{35}S-labelled BP1 decreases significantly compared to the other proteasome proteins in just 30 mins and is thought to be an *in vivo* assembly intermediate[1]. Several SILAC experiments were performed to determine the rate of turnover of the various proteins that comprise the BP1 complex.

The traditional method for measuring protein degradation is to use a 'pulse-chase' type of experiment where cells are 'pulse'-labelled with radioactive elements such as ^{35}S which is present in the growth media. Following sufficient labelling, the cells are then 'chased' with normal media containing no radioactive elements. By collecting cell lysate at periodic time points, isolating the protein of interest (immuno- or affinity tag purification) and determining the amount of radioactivity (scintillation counting) over time, it is possible to determine the rate of degradation of a particular protein. During the 'chase' part of the experiment, the radioactive-labelled proteins are degraded and replaced with proteins that do not contain radioactive elements and, therefore, are not included in the radioactive measurements.

An alternative approach is to use SILAC where, instead of the 'pulse-labelling' being with radioactive ^{35}S, stable isotope labels are used. By measuring the ratio of heavy to light peptides, the protein flux can be determined. This is slightly different from measuring the degradation rate using the radioactive pulse-chase since in the 'chase' part of the experiment as the cells grow, newly synthesised proteins effectively 'dilute' the original 'pulse'-labelled proteins, and therefore the protein flux is measured rather than the protein degradation rate. However, by considering the growth rate of the cells, an estimate of the protein degradation can be obtained as described[2].

My contribution to the project was to determine the rate of incorporation of the four proteins into the BP1 complex, and involved a small modification of the classic SILAC protocol. HeLa lysate was run on a BN gel to preserve complexes. The rate of incorporation of heavy leucine in the BP1 complex, as well as in the full proteasome complex, was determined for the four proteins of interest (PSMC1, PSMC2, PSMD2 and PSMD5), the percentage incorporation of 'heavy' leucine over time being determined. HeLa cells were grown in 'light media' (containing naturally- light leucine) and were switched over to 'heavy media' (containing 'heavy' leucine) at the start of the time course. Cells were harvested and lysed at different time points. By measuring the heavy/light ratio of the leucine containing peptides, an indication of protein degradation rate and turnover was obtained. It would be expected that the proteins in the BP1 complex would have a higher turnover rate than the proteins in the proteasome.

In total, three modified SILAC experiments were performed. The first SILAC experiment used time points between 0 and 150 min (0, 5, 15, 30, 60 and 150 mins) but unfortunately very little heavy leucine incorporation was observed. The second SILAC experiment used time points between 0 and

1440 min (0, 60, 105, 150, 195 and 1440 mins) but heavy leucine incorporation was only observed for the later time points. The third SILAC experiment used time points of 0 to 48 hours (0, 4, 8, 12, 18, 24 and 48 hours). It was originally thought that the turnover of the BP1 complex would be fairly fast, hence the short time frames chosen initially. However, the delayed uptake of heavy amino acids might explain the slower dynamics[2]. Various rationales have been postulated to explain this delay in the uptake of heavy amino acids, although the exact reasons for this are not known. It could be due to amino acid recycling, since if the intracellular stores of light leucine are significantly high at the start of the 'chase' phase this could result in a 'blurred' transition between the 'pulse' and 'chase' phases. Another reason could be that some nascent proteins are more unstable than the more mature equivalent proteins which are already in complexes (and which contain light leucine). This has, for example, been shown for proteins that have problematic tertiary folding[3-5]. Other cellular functions that could explain the delay in uptake and the delay in synthesis include mRNA transport and regulated translation intitiation[6-7]. However, since the HeLa cells were asynchronous, these factors should not have had a significant effect.

The main challenge with this experiment was data extraction and quantification. The data were initially quantitated using Vista, an in-house software suite for quantitation of isotopically labelled proteins[2, 8]. Another challenge was that the proteins comprising the BP1 complex were not particularly abundant and therefore it was not always possible to identify them for all the time points. Even if a protein is identified at every time point, the particular peptides detected might be different. It is not ideal, and it is actually invalid to assess protein degradation from the labelling of different component peptides across the different fractions. To improve detection of all the BP1 complex proteins at all time points, an inclusion list of the peptide m/z values (2+, 3+) was generated and the samples run with the inclusion list to try to ensure that the same peptides were identified in each fraction. Despite identifying many more peptides, all the time points were still not covered.

A solution to the data analysis problems was to use an alternative method of analysis: to quantitate using a specific heavy/light peptide using the peptide EIC in Quanbrowser (Xcalibur), similar to the method used for KAYAK quantitation, rather than relying peptide detection by MSMS. This manual approach was time consuming and therefore only one/two peptides per protein were quantitated. However, it meant that the proteins could be quantitated at each time point and the initial results were very promising (Figure A.2). The percentage incorporation of heavy leucine was determined for PSMC1 and PSMC2 in both the BP1 complex (gel slice 5: PSMC1_5, PSMC1_5) as well as in the 26S proteasome (gel slice 2: PSMC1_2, PSMC1_2). The rate of heavy leucine incorporation was much faster for the proteins in the BP1 assembly complex compared to the 26S proteasome, suggesting that the formation of the BP1 complex is much faster than that of the 26S proteasome complex. PSMD5 was only found in the BP1 complex and not in the 26S proteasome, which is consistent with the hypothesised molecular chaperone function of PSMD5. The rate of heavy leucine incorporation into PSMD5 is similar to that of the proteins in the 26S proteasomes, which was expected, since PSMD5 is a molecular chaperone and is 'recycled' and does not end up in the final

26S complex. Therefore, PSMD5 was expected to have a lower heavy leucine incorporation rate than the other constituents of the BP1 complex (that end up finally in the 26S proteasome) and for this reason would need to have a higher synthesis rate. This is consistent with the model for base assembly as described[1, 9].

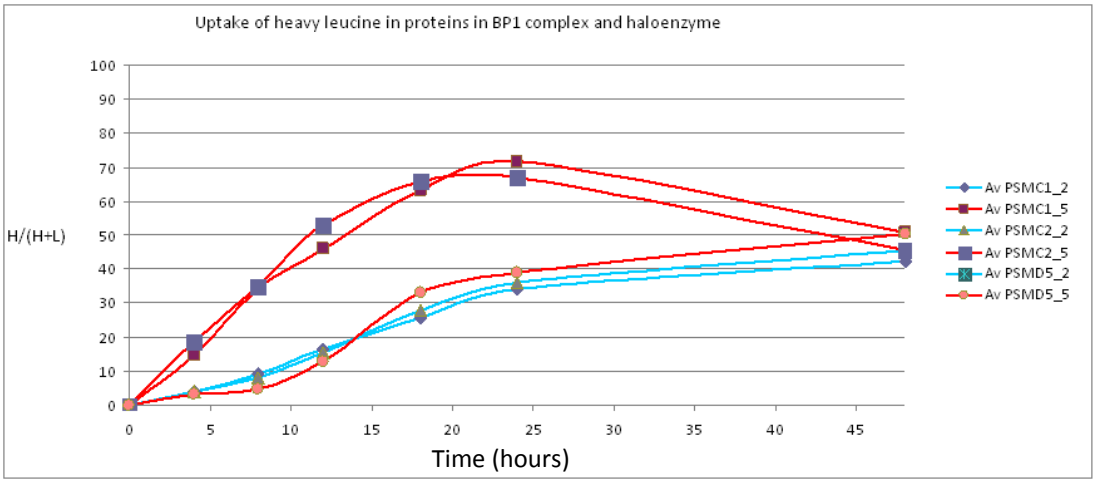


Figure A.2 Percentage incorporation of heavy leucine over a period of 48 hours. Av PSMC1_2 indicates the average (of 2 peptides) for protein PSMC1 in gel slice 2 which comprises 26S proteasome, whereas PSMC1_5 indicates the same protein in gel slice 5 which comprises the BP1 complex.

A.2 Prokaryotic Ubiquitin-Like Protein (Pup) Proteome of *Mycobacterium tuberculosis*

OPEN ACCESS Freely available online PLoS one

Prokaryotic Ubiquitin-Like Protein (Pup) Proteome of *Mycobacterium tuberculosis*

Richard A. Festa[1☯], Fiona McAllister[2☯], Michael J. Pearce[1☯], Julian Mintseris[2], Kristin E. Burns[1], Steven P. Gygi[2*], K. Heran Darwin[1*]

1 Department of Microbiology, New York University School of Medicine, New York, New York, United States of America, **2** Department of Cell Biology, Harvard Medical School, Boston, Massachusetts, United States of America

Abstract

Prokaryotic ubiquitin-like protein (Pup) in *Mycobacterium tuberculosis* (*Mtb*) is the first known post-translational small protein modifier in prokaryotes, and targets several proteins for degradation by a bacterial proteasome in a manner akin to ubiquitin (Ub) mediated proteolysis in eukaryotes. To determine the extent of pupylation in *Mtb*, we used tandem affinity purification to identify its ''pupylome''. Mass spectrometry identified 55 out of 604 purified proteins with confirmed pupylation sites. Forty-four proteins, including those with and without identified pupylation sites, were tested as substrates of proteolysis in *Mtb*. Under steady state conditions, the majority of the test proteins did not accumulate in degradation mutants, suggesting not all targets of pupylation are necessarily substrates of the proteasome under steady state conditions. Four proteins implicated in *Mtb* pathogenesis, Icl (isocitrate lyase), Ino1 (inositol-1-phosphate synthase), MtrA (*Mtb* response regulator A) and PhoP (phosphate response regulator P), showed altered levels in degradation defective *Mtb*. Icl, Ino1 and MtrA accumulated in *Mtb* degradation mutants, suggesting these proteins are targeted to the proteasome. Unexpectedly, PhoP was present in wild type *Mtb* but undetectable in the degradation mutants. Taken together, these data demonstrate that pupylation regulates numerous proteins in *Mtb* and may not always lead to degradation.

Citation: Festa RA, McAllister F, Pearce MJ, Mintseris J, Burns KE, et al. (2010) Prokaryotic Ubiquitin-Like Protein (Pup) Proteome of *Mycobacterium tuberculosis*. PLoS ONE 5(1): e8589. doi:10.1371/journal.pone.0008589

Editor: Anil Kumar Tyagi, University of Delhi, India

Received November 17, 2009; **Accepted** December 9, 2009; **Published** January 6, 2010

Funding: This work was supported by National Institutes of Health (NIH) grants GM67945, HG3456, and HG3616 awarded to S.P.G., and AI065437, HL092774 and a Center for AIDS Research Pilot Project grant (A1027742-17) awarded to K.H.D. M.J.P. received support from 5T32AI07189-25 and R.A.F. was the recipient of a Jan T. Vilcek Fellowship. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: steven_gygi@hms.harvard.edu (SPG); heran.darwin@med.nyu.edu (KHD)

☯ These authors contributed equally to this work.

As part of a collaboration with Professor Heran Darwin at the New York University School of Medicine, I was involved in identifying which protein sites had been modified with pup, a postranslation modification. My contribution included sample preparation and mass spectrometric analysis using GeLCMSMS[10].

A.3 Selective Chemical Intervention in the Proteome of Caenorhabditis elegans

Selective Chemical Intervention in the Proteome of *Caenorhabditis elegans*

Holger Husi[1], Fiona McAllister[2], Nicos Angelopoulos[1], Victoria J. Butler[3], Kevin R. Bailey[4], Kirk Malone[4], Logan MacKay[2], Paul Taylor[1], Antony P. Page[3], Nicholas J. Turner[4], Perdita E Barran[1,2**] and Malcolm Walkinshaw[1**]

**Authors to whom correspondence should be addressed
[1] Centre for Translational and Chemical Biology, University of Edinburgh, Edinburgh EH9 3JJ UK
[2] School of Chemistry, University of Edinburgh, Edinburgh EH9 3JJ UK
[3] Faculty of Veterinary Medicine, University of Glasgow, 464 Bearsden Road, Glasgow G61 1QH UK
[4] School of Chemistry, University of Manchester, 131 Princess Street, Manchester, M1 7DN, UK

Abstract

We present the first study of protein regulation by ligands in *C.elegans*. The ligands were peptidyl-prolyl isomerase inhibitors of cyclophilins. Up-regulation is observed for several heat shock proteins and one ligand in particular caused a greater than two-fold enhancement of cyclophilin CYN-5. Additionally several metabolic enzymes display elevated levels. This approach, using label free relative quantification, provides an extremely attractive approach to the effect of ligands on an entire proteome, with minimal sample pre-treatment, which could be applicable to large- scale studies. In this initial study, which compares the effect of three ligands, 54 unique proteins have been identified that are up (51) or down (3) regulated in the presence of a given ligand. A total of 431 C. *elegans* proteins were identified. Our methodology provides an intriguing new direction for *in-vivo* screening of the effects of novel and untested ligands at the whole organism level.

I was involved in carrying out the mass spectrometry work for the LCMS label free comparison and the LCMSMS identification[11].

B. LTQ-Orbitrap

The Orbitrap mass analyser was developed by Alexander Makarov and it is a modified version of the ion trap that Kingdon developed in the 1920s.[12-13] An Orbitrap is different to a conventional ion trap in that the ions are trapped in an electrostatic field: there is no magnet or RF used to trap ions. The Kingdon trap comprises a thin wire central electrode, an outer coaxilal cyclindrical electrode along with two endcap electrodes. Similar to the Kingdon trap, the Orbitrap comprises inner and outer coaxial electrodes. However, in the Orbitrap, the inner electrode is spindle-shaped.[14] In the Orbitrap, tangentially injected ions are attracted electrostatically to the inner electrode but this is balanced by the centrifugal force of the ions. The electrostatic field inside the Orbitrap causes the ions to cycle the central electrodes where the axial frequency of these oscillations is inversely proportional to the square root of the m/z ratio and is independent on the energy or spatial spread of the ions.[13] Similar to in FTICR, the trap can be used as a mass analyser. The Orbitrap is surrounded by an outer electrode that is split in the middle by an insulating ring. The ions induce an image current and this is detected by a differential amplifier between the two sides of the outer electrode surrounding the Orbitrap.[15] The m/z ratio of the ions is determined by a fast fourier transformation (FFT) carried out on the oscillating frequencies.

The LTQ Orbitrap hybrid mass spectrometer (Figure B.1) is ideally suited to 'bottom up' proteomics. Both mass analysers can work in parallel: while the high resolution/high mass accuracy MS scan of the precursor ions is acquired in the Orbitrap mass analyser, many fast MSMS spectra can be obtained on the lower resolution/lower mass accuracy LTQ mass analyser. The parallel operation of both the LTQ and the Orbitrap is achieved as the initial part of the high resolution spectra is used for the data-dependent precursor selection which are subsequently fragmented while the high resolution Orbitrap spectrum is being acquired.

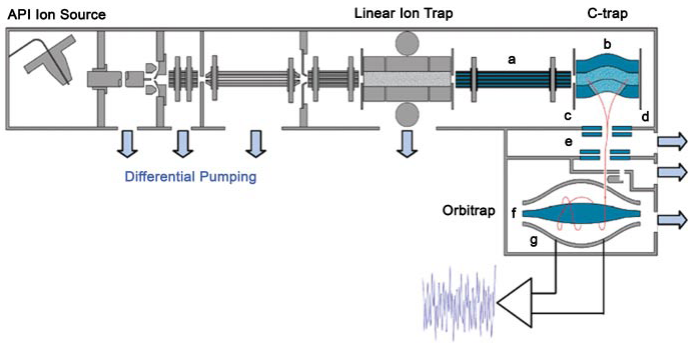


Figure B.1 Schematic of the LTQ-Orbitrap Discovery which couples an ESI source to a linear ion trap and an Orbitrap. (a) transfer octopole; (b) curved RF-only quadrupole (C-trap); (c) gate electrode; (d) trap electrode; (e) ion optics; (f) inner Orbitrap electrode (central electrode); (g) outer Orbitrap electrode. Adapted from Perry *et al*.[13]

C.Peptide Library approach to identify peptide substrates

C.1 Phosphorylation site determination for peptide libraries

Three peptide libraries were used: acidic, basophilic and proline directed (Table C.1). A modified KAYAK assay was performed on three peptide libraries. The acidic library and the proline directed library were incubated with 3 different cell lysates: nocodazole arrested HeLa, and HEK 293 treated with EGF and treated with PMA. The basophilic library was incubated with 3 different cell lysates: MCF7, and HEK 293 treated with EGF and treated with PMA. Details of lysate preparation are provided in the Section 2.7.2 For each KAYAK assay 20 µg of cell lysate was, individual concentrations of the peptide libraries were approximately 2.6 µM, 1.5 µM and 3.2 µM for the acidic, basophilic and proline directed libraries, respectively. The KAYAK protocol used for the peptide library assays is the same as described in Section 2.6.2.1 with the following exception: no internal standard ('heavy library') was added. Internal standard was not added since the focus of the first experiment was to determine which sites are phosphorylated rather than their quantitation. The peptides were analysed by LC-MS/MS (LTQ-Orbitrap) (details in Experimetal Procedures Section 2.5) and searched using the SEQUEST algorithm [16]against a reversed database containing all the sequences in the libraries and a concatenated *E. coli* database. Peptides were filtered to a peptide FPR < 0.1%. The phosphorylated peptides for each library were submitted to MotifX[17] to generate phosphorylation motifs.

Table C.1 Sequences in the acidic, basophilic and proline directed peptide libraries. The heavy isotope version contains a heavy leucine (shown in lower case) and a phosphorylated S and/or T (shown in lower case).

	Peptides in library	Sequences in the libraries
Acidic	864	L[LE]D[KDN]D[DA][LE][ST]D[EL]E[LEN][EL]K l[LE]D[KDN]D[DA][LE][st]D[EL]E[LEN][EL]K (heavy isotope)
Basophilic	576	[NKP][LK][KA]R[KRS]L[ST][SA][EPL][DL]LAA [NKP][LK][KA]R[KRS]l[st][SA][EPL][DL]LAA (heavy isotope)
Proline directed	576	[KP]L[VKE]L[AP][NE][ST]P[KI][LKP]VV[KL] [KP]L[VKE]l[AP][NE][st]P[KI][LKP]VV[KL] (heavy isotope)

C.2 Relative KAYAK using peptide libraries

A relative KAYAK assay was performed using the three libraries and HEK 293 cell lysate with and without EGF treatment. The same procedure as in Section C.1 above was performed except that following incubation of the libraries with the lysate, the appropriate internal standard ('heavy library') was added to the reaction mix following quenching. Following database searching, the heavy to light ratios were quantified using Vista (an in-house quantitation software suite) and the amount of phosphorylation was compared between the three libraries with and without EGF treatment.

C.3 Peptide library approach to identify substrates for KAYAK

For kinase activity profiling using KAYAK it is critical to have good peptide substrates. Designing good peptide substrates is not straightforward. One approach that can be used to identify peptide substrates is to use peptide libraries which are generated using combinatorial peptide chemistry.

Three different combinatorial libraries were investigated: acidic, basophilic and proline directed (Table C.1), to screen for interesting peptide substrates for KAYAK. The first experiment involved incubating the libraries with various different lysates and determining how many of the peptides were phosphorylated. In total, 50, 218 and 134 phosphorylated peptides were observed for the acidic, basophilic and proline directed libraries, respectively (Table C.2). Motif consensus software was used (Motif-X[17]) to generate putative phosphorylation motifs for each of the libraries using the phosphorylated peptide sequences observed.

Table C.2 Table showing number of phosphorylated substrate peptides for each of the samples.

Library	Total peptides in library	Cell lysate	Substrate conc (uM)	Phosphopep per sample	Phosphopep per library
Acidic	864	293T – PMA stimulated	2.57	27	54
		HeLa – nocodazole arrested		8	
		293T – EGF stimulated		30	
Basophilic	576	293T – EGF stimulated	1.46	221	323
		MCF7		114	
		293T – PMA stimulated		116	
Proline directed	576	293T – EGF stimulated	3.16	54	121
		HeLa – nocodazole arrested		93	
		293T – PMA stimulated		38	

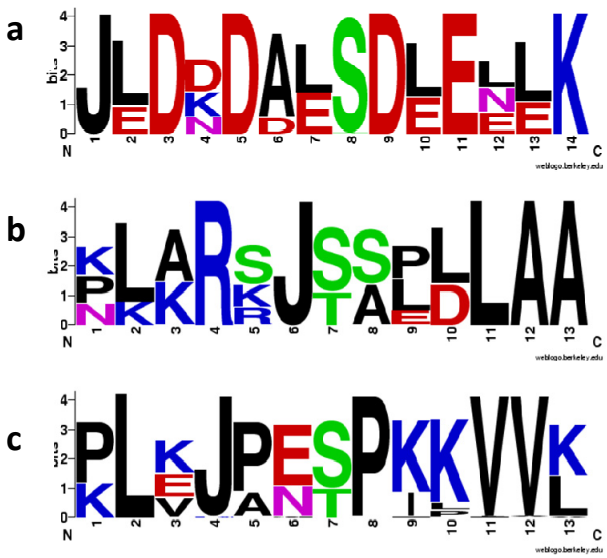


Figure C.1 Motif for (a) acidic library, (b) basophilic library, (c) proline directed library. Colour code for amino acids: Green - polar, blue – basic, black – hydrophobic amino acids. Motifs generated using MotifX[17].

For the acidic library 50 peptides were phosphorylated, which is of particular interest since it is unusual to observe the acidic peptides being phosphorylated. Only one acidic kinase is currently known (casein kinase).

A relative KAYAK experiment was carried out comparing the change in phosphorylation of the library peptides in two different samples: cells (293T) with and without EGF stimulation. Substrates displaying a change in the degree of phosphorylation between cells treated in different conditions are desired. One of the difficulties with the relative KAYAK quantitation was that many of the peaks had the same or similar m/z in the same time window due to the nature of the peptide sequences in the library. The peptides that showed a significant degree of phosphorylation (light/heavy ratio), and where the degree of phosphorylation was different between samples with and without EGF treatment, were selected and synthesised by Cell Signalling Technologies (Table C.3)

Table C.3 Peptide sequences selected from the libraries for synthesis.

Library	Sequence	L/H ratio, EGF treatment	L/H ratio, no EGF treatment
Acidic			
pept_lib_1_57	JLDDDALsDLELEK	1.4	0
pept_lib_1_552	JEDNDAEsDLENLK	0.89	0
Basophilic			
pept_lib_2_104	KLARSJsAEDLAA	0	0.2
pept_lib_2_594	PKKRRJsSPLLAA	0	0.2
pept_lib_2_657	PLARKJsAPLLAA	0	0.09
SP Directed			
pept_lib_3_117	KLKJPEsPKKVVK	0	1.4
pept_lib_3_141	KLKJPEtPKKVVK	0.2	0.4
pept_lib_3_417	KLKJPNtPKKVVL	1	1.6

D. Quantitation using targeted spectral counting

Comparison of normal and targeted spectral counting approach for kinase quantitation with spectral counting.

Table D.1 List of kinases and their spectral counts using a normal and targeted spectral count approach. A list of m/z kinase peptides (from previous analyses) was used as an inclusion list. Three gel bands (B, C, D) were analysed using target approach (shown in grey) and a TOP 10 data (shown in white) dependent acquisition strategy. A greater number of kinases and spectral counts were observed using the target approach.

Gene symbol	MB_Target	MB_Top10	MC_Target	MC_Top10	MD_Target	MD_Top10	MBCD_Target	MBCD_Top10	Target/Top10
CDC2	44	4	64	5	63	4	171	13	13.2
CDK2	1	0	1	0	2	0	4	0	#DIV/0!
CDK3	4	0	23	2	16	2	43	4	10.8
CDK4	6	1	6	1	5	1	17	3	5.7
CDK9	4	0	0	0	0	0	4	0	#DIV/0!
CSK	0	0	0	0	1	1	1	1	1.0
CSNK1A1L	8	1	9	1	4	0	21	2	10.5
CSNK2A1	51	11	43	5	0	0	94	16	5.9
CSNK2A2	50	5	24	2	0	0	74	7	10.6
EGFR	0	0	23	3	0	0	23	3	7.7
EPHA2	0	0	1	0	5	0	6	0	#DIV/0!
ERBB4	0	1	0	1	0	0	0	2	0.0
FGFR1	6	1	6	1	5	1	17	3	5.7
GSK3B	1	0	1	0	0	0	2	0	#DIV/0!
ILK	3	0	0	0	0	0	3	0	#DIV/0!
MAP2K1	0	0	0	0	11	1	11	1	11.0
MAP2K2	0	0	1	0	12	1	13	1	13.0
MAP2K3	0	0	0	0	1	0	1	0	#DIV/0!
MAP3K1	0	0	0	0	1	0	1	0	#DIV/0!
MAPK1	0	0	1	1	25	1	26	2	13.0
MAPK15	0	0	0	0	3	0	3	0	#DIV/0!
MAPKAPK2	2	0	3	1	0	0	5	1	5.0
MAPKAPK3	0	0	0	0	1	0	1	0	#DIV/0!
MASTL	0	0	2	0	0	0	2	0	#DIV/0!
NEK11	0	1	0	0	0	0	0	1	0.0
NEK7	1	0	3	0	1	0	5	0	#DIV/0!
NRBP1	7	1	1	0	0	0	8	1	8.0
OXSR1	1	0	31	4	92	9	124	13	9.5
PAK1	0	0	0	0	5	0	5	0	#DIV/0!
PAK2	0	0	0	0	45	3	45	3	15.0
PAK3	0	0	1	0	18	2	19	2	9.5
PBK	0	0	7	0	41	3	48	3	16.0
PIK3R4	0	1	0	0	0	0	0	1	0.0
PKN2	1	0	63	8	4	0	68	8	8.5
PLK1	20	2	4	5	0	0	24	7	3.4
PRKACA	0	0	8	1	34	4	42	5	8.4
PRKACB	0	0	0	0	4	1	4	1	4.0
PRKACG	0	0	0	0	1	0	1	0	#DIV/0!
PRKDC	511	101	1	0	0	1	512	102	5.0
RIOK1	1	0	5	0	0	0	6	0	#DIV/0!
RIOK2	0	0	4	0	0	0	4	0	#DIV/0!
RIPK2	0	0	0	0	1	0	1	0	#DIV/0!
ROCK1	2	3	0	0	0	0	2	3	0.7

ROCK2	52	9	0	0	0	0	52	9	5.8
RPS6KA3	0	0	0	0	9	0	9	0	#DIV/0!
RPS6KA6	0	0	0	0	0	1	0	1	0.0
SCYL1	2	1	0	0	0	0	2	1	2.0
SCYL2	0	0	1	0	0	0	1	0	#DIV/0!
SLK	22	10	0	0	0	0	22	10	2.2
SRPK1	1	0	2	1	0	0	3	1	3.0
STK10	0	0	1	0	0	0	1	0	#DIV/0!
STK24	0	0	13	1	33	4	46	5	9.2
STK25	0	0	0	0	4	1	4	1	4.0
TGFBR1	1	0	7	0	6	0	14	0	#DIV/0!
TRIM28	318	43	54	6	11	1	383	50	7.7
TRIM33	0	1	0	0	0	0	0	1	0.0
TRRAP	0	0	0	0	0	1	0	1	0.0
TTBK2	5	1	0	0	0	0	5	1	5.0
TTK	0	0	2	0	0	0	2	0	#DIV/0!
TTN	0	3	2	1	0	0	2	4	0.5

a

b

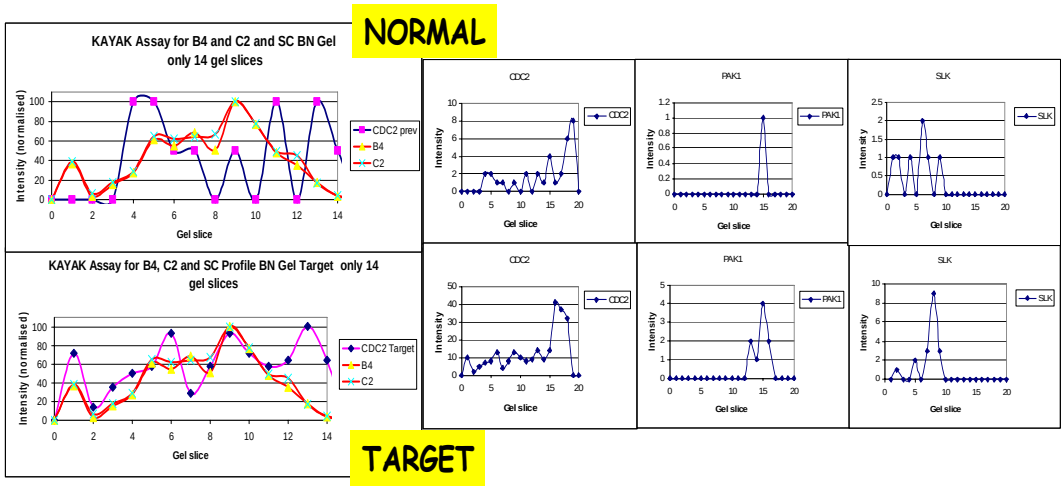


Figure D.1 Comparison between normal and target spectral counting quantitation approaches. The protein correlation profiles of the peptides C2 and B4 (known to be substrates of CDC2) with CDC2 are shown with (a) 'normal' spectral counting and (b) spectral counting using an inclusion list. Better correlation is observed for the target approach.

E. Label free software packages

E.1 XCMS

XCMS was written using R, which is an open source statistical programming environment. Flowcharts showing the XCMS processing and preprocessing strategy are illustrated in Figure E.1[18] .

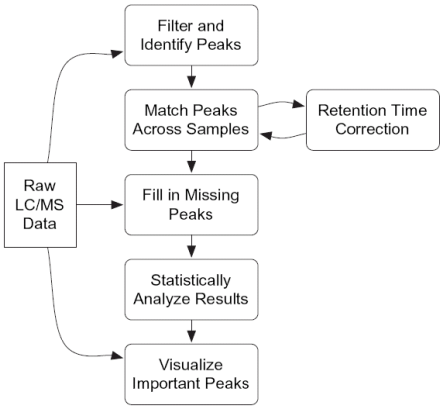


Figure E.1 Overview of XCMS processing[18].

E.1.1 Peak detection

Since many peaks may be split between two bins, adjacent chromatographic slices are combined. Following generation of the combined chromatograms a matched filtrate algorithm is employed using a second-derivative Gaussian as the model peak shape[18]. Whilst the boundaries for peak integration are determined using the zero-crossing points of the filtered chromatograms, it is the unfiltered chromatograms that are actually integrated. Where the data resolution is very high and the m/z peak width is narrow compared to the 0.1 m/z bin size or where the data is centroided where the peak width is infinitely thin, the maximum signal intensity from adjacent slices can be overlapped. Where the resolution is low and the m/z peak width is large compared to the 0.1 m/z bin, XCMS's vicinity elimination post processing step removes peaks close to the higher intensity peak within a 0.7 m/z range.

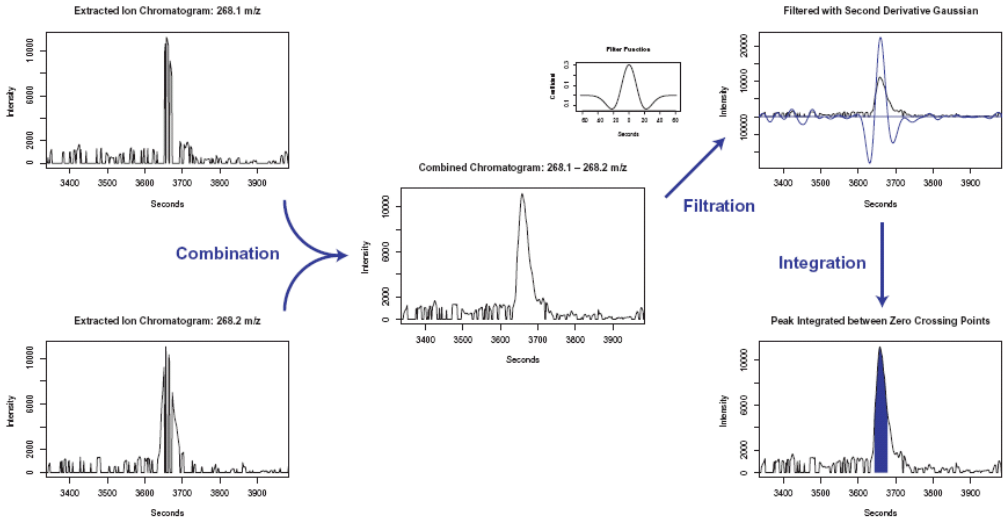


Figure E.2 LC/MS filtering and peak detection in XCMS[19].

E.1.2 Peak matching

To overcome the problem of splitting due to arbitrary bin boundaries, peaks from overlapping bins are later removed in the postprocessing operations. Following mass binning (mass bins of width 0.25 m/z are used), the peak matching algorithm groups peaks from different samples into each bin by calculating the overall distribution of peaks in chromatographic time. This is determined using the kernel density estimation method which is effectively creating a smoothed histogram by applying a Gaussian function. An example of cross-sample matching is illustrated in Figure E.3.

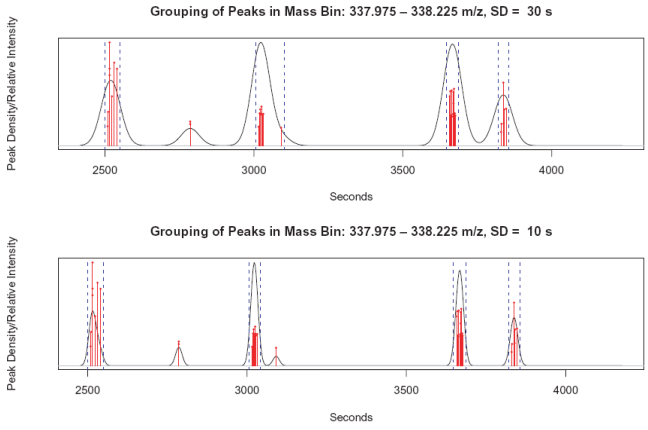


Figure E.3 Cross sample peak matching in XCMS.

E.1.3 Retention time alignment

If there are regions of the spectra where no well behaved peak groups exist, XCMS uses an algorithm to estimate the retention time deviation. For those peak groups which are missing a sample, XCMS adds in intensity values by using integration data from the raw LCMS runs. XCMS generates EIC for significant peaks where the alignment can be observed.

Work is currently in progress to standardise LCMS data to be able to combine and compare LCMS data obtained from different instrument platforms. The Annotated Putative Peptide Markup Language (APML)[20] along with the Corra framework[20] are such initiatives are currently under development.

E.2 MS-Xelerator

The software from MsMetrix is user-friendly and comparison can be performed either manually or using algorithmic techniques to identify changes. Manual comparisons can be carried out by comparing the TCMs (Total Chromatogram Mass spectra), which are the average spectra of an LC-MS run, or by comparing the TIC (Total Ion Chromatograms) for specific mass ranges. The advantage of using 'mass spectra' is that alignment problems are minimal. The disadvantage is that it is difficult to detect small differential peaks in the presence of larger peaks having the same or similar m/z values. It is also possible to set up automatic detection of differences. Another method is to scan all the 2D-LC-MS surfaces simultaneously, creating a binary map of m/z vs. retention time and then unique peaks can be identified.

In MSX it is possible to normalise the spectra based on area, a unit vector or a specific peak. The disadvantage of using normalisation using the area of the TIC or that of a single peptides is that it is a single measurement and therefore may not be reliable. Also, if sample comparison is based on total protein abundance, then it is not accurate to correlate the TIC with total protein since there are likely to be other compounds present such as lipids and small molecules that make up the TIC.

E.3 Progenesis

Nonlinear Dynamics (UK) They offer a series of packages for biomarker discovery for different instrument platforms: 1D/2D gel quantitation, MALDI and LCMS. and it is particularly useful to visualise the error bars for the peak abundance both in and between the groups. The peptides are ranked by ANOVA p-value and PCA is possible.

Following quantitation of the EIC from the LCMS trace it is possible to export the significant peaks into a file that can be exported to Mascot (or Sequest or Phenyx). Following database searching, the Mascot results can be exported into xmL format and imported back into Progenesis and the peptides are automatically matched to the MSMS spectra and protein identifications. Where one peptide

matches several proteins, all the possible protein identifications are listed. One of the advantages of Progenesis is that the link between LCMS and LCMSMS is not restricted to a single LCMSMS run, or indeed to the same instrument. LCMS experiments can be performed on an FTICR mass spectrometer and quantitatively processed whilst LCMSMS can be performed on an ion trap. Following alignment and quantitation of the FTICR LCMS, the LCMSMS spectra can be aligned 'on top' of the FTICR LCMS data and then processed to obtain a peak list to submit to Mascot.

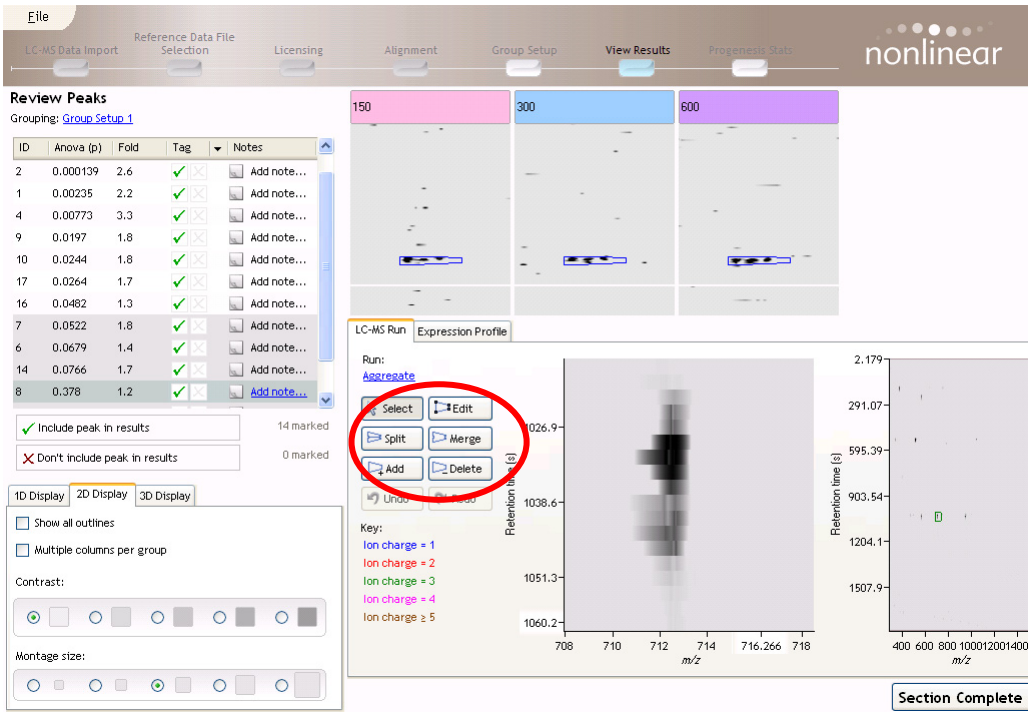


Figure E.4 Screenshot of analysis of three different peptide concentrations using Progenesis. Progenesis allows the user to change the area of the peak and the isotopes included (indicated by the red circle).

F. Protein complexes in saliva

F.1 BN PAGE of saliva

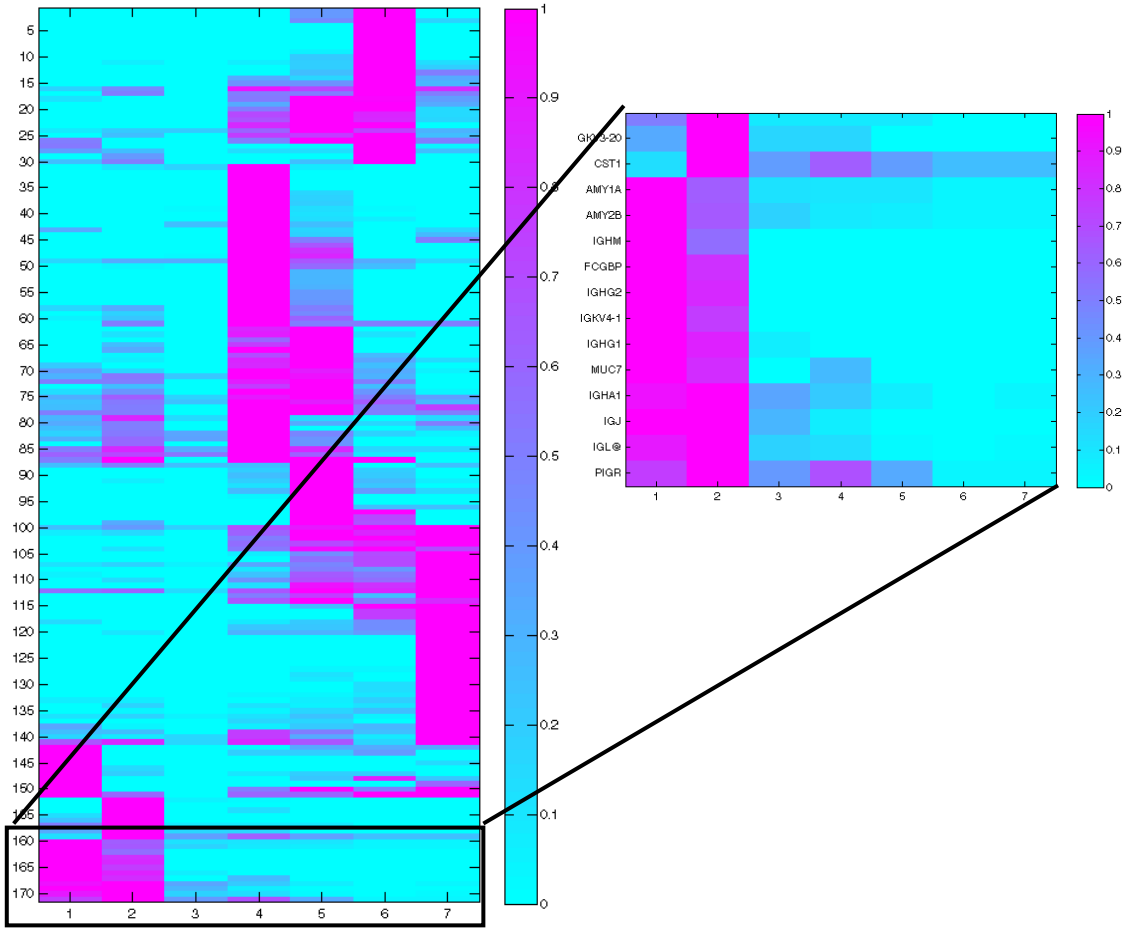


Figure H.1 Profile of salivary protein complexes. Saliva was run on a Blue native PAGE gel and seven slices were analysed by GeLCMSMS. Spectral counts of the proteins give an estimate of the protein abundance in each of the fractions. (a) Profile of protein complexes for each gel slice. Hierarchical clustering grouped proteins with similar spectral count profiles together. Proteins were normalised to the gel slice with the highest number of spectral counts. Only proteins with a total spectral greater than 5 were considered. The enlarged snapshot of (a) shows putative salivary protein complex comprising amylase and Igs with a smiliar composition to that in Figure 3.8.

F.2 SDS PAGE of saliva

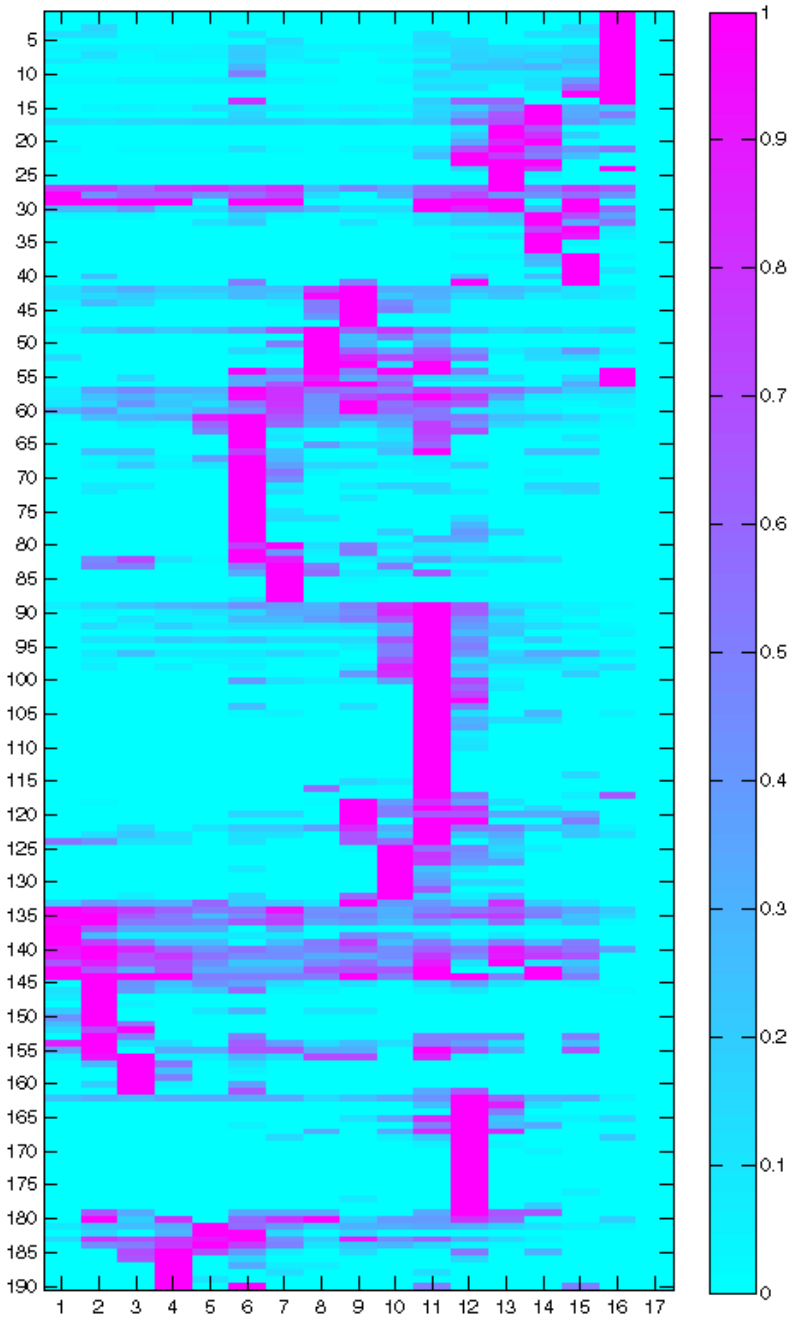


Figure H.2 Profile of salivary protein complexes. Saliva was run on an SDS-PAGE gel and 17 slices were analysed by GeLCMSMS. Spectral counts of the proteins give an estimate of the protein abundance in each of the fractions. Hierarchical clustering grouped proteins with similar spectral count profiles together. Proteins were normalised to the gel slice with the highest number of spectral counts. Only proteins with a total spectral greater than 5 were considered.

F.3 BN PAGE of saliva residue and tear supernatant

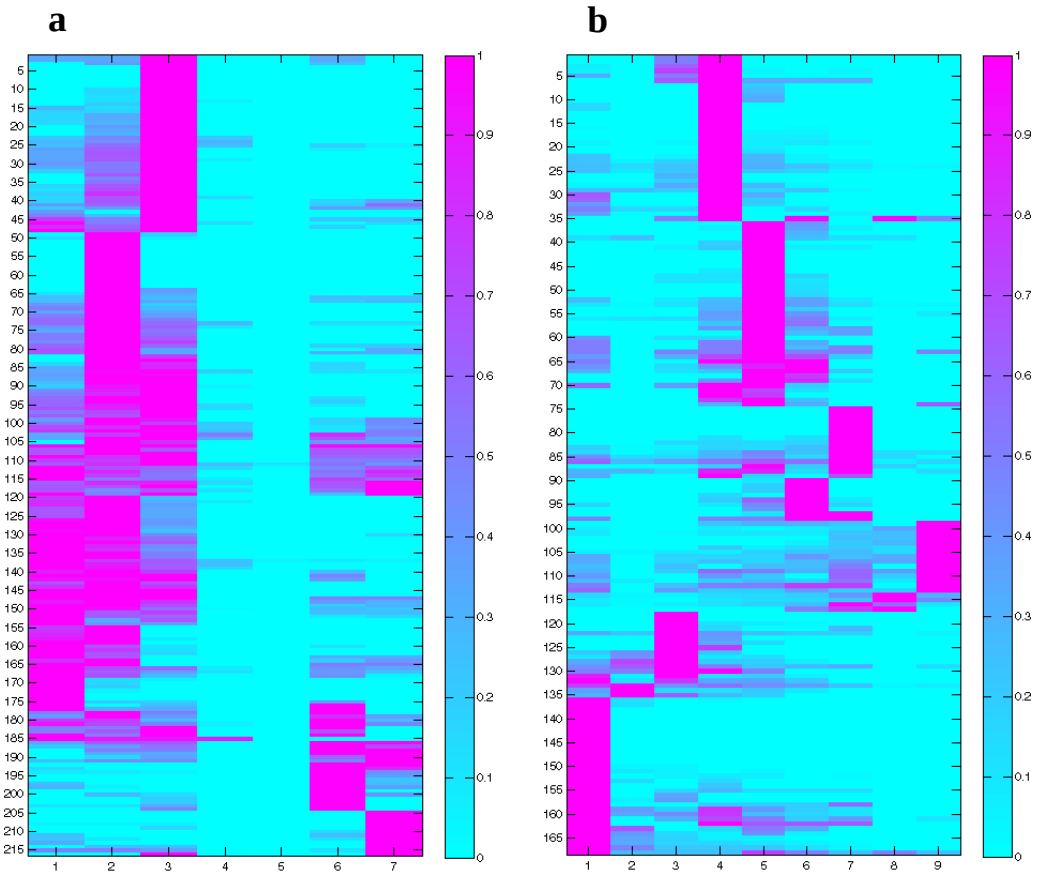


Figure H.3 Profile of protein complexes. (a) Saliva residue run on BN gel and divided into 7 slices and analysed by GeLCMSMS. (b) Tear supernatant run on BN gel and divided into 10 fractions which were analysed using GeLCMSMS. Spectral counts of the proteins give an estimate of the protein abundance in each of the fractions. Hierarchical clustering grouped proteins with similar spectral count profiles together. Proteins were normalised to the gel slice with the highest number of spectral counts. Only proteins with a total spectral greater than 5 were considered.

G. Kinase correlation profiling of saliva on BN PAGE

Kinase activity profiles (obtained using KAYAK) were correlated with kinase abundance profiles (obtained from GeLC-MSMS).

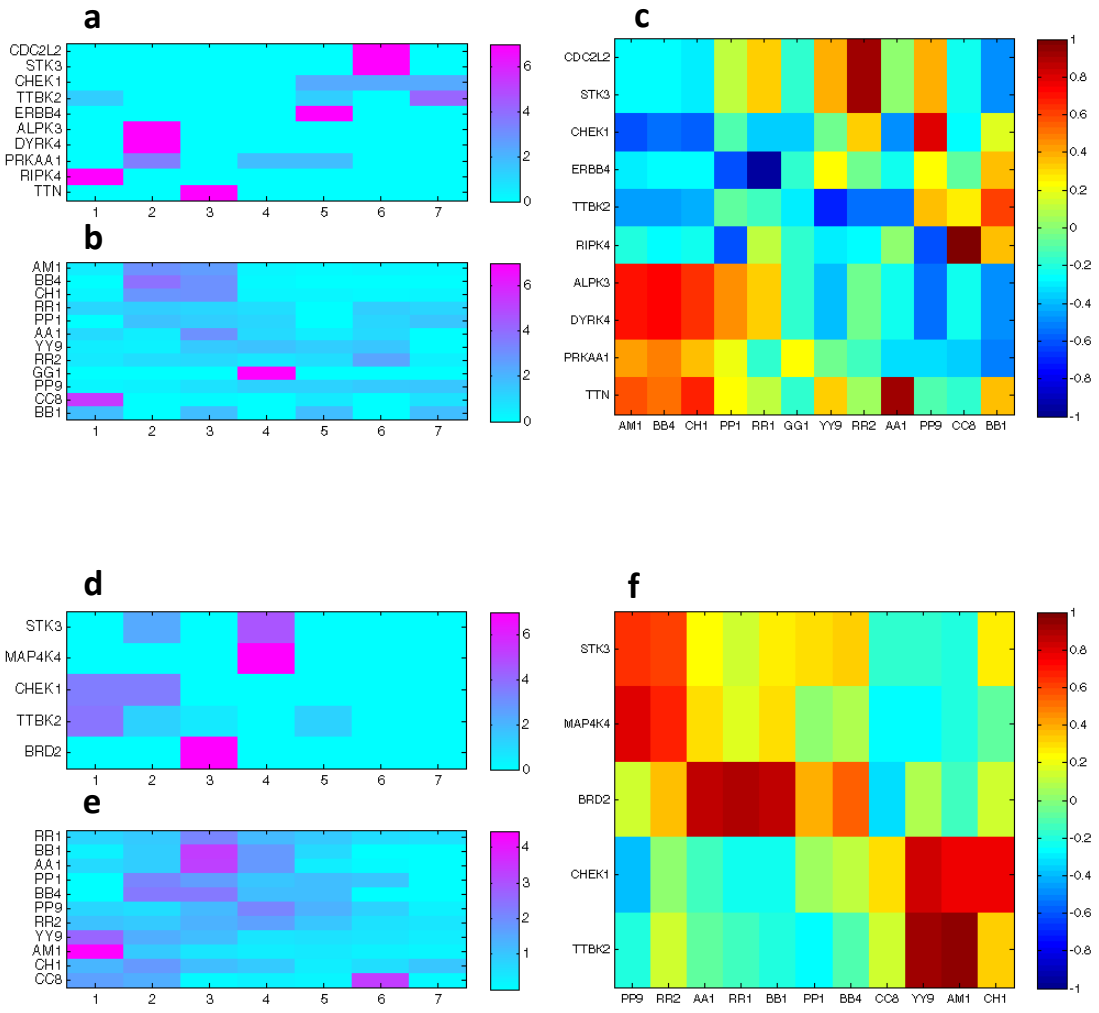


Figure F.1 BN PAGE KAYAK correlation profiling on saliva supernatant (a-c) and saliva residue (d-f). (a) Spectral count profile of kinases in the saliva supernatant. Normalisation to the sum of the spectral counts in each row. (b) Kinase acitivity (KAYAK) of saliva supernatant. Normalisation to the sum of the kinase activites in each row (c) Heatmap of Pearson correlation coefficients bewteen kinase abundance (SC) and activity (KAYAK) in saliva supernatant. (d) Spectral count profile of kinases in the saliva supernatant. Normalisation to the sum of the spectral counts in each row. (e) Kinase acitivity (KAYAK) of saliva supernatant. Normalisation to the sum of the kinase activites in each row (f) Heatmap of Pearson correlation coefficients bewteen kinase abundance (SC) and activity (KAYAK) in saliva supernatant.

H. Optimisation of parameters for label free software packages

There were various label-free software packages available at the time of analysis, both open source and commercial. Three of these software packages were investigated: MS-Xelerator, Progenesis and XCMS. Details of optimisation are shown below.

Table G.1 Optimisation of MSX minimum peak width setting for 20 minute peptide run on HCT.(300/150 fmol)

Minimum peak width (mins)	Number of peaks	Average	SD
0.1	111	2.61	1.52
0.2	92	2.63	1.62
0.3	57	2.26	0.46
0.4	34	2.16	0.40
0.5	26	2.07	0.28
0.6	18	2.00	0.22
0.7	9	1.99	0.21
0.8	6	1.98	0.24
0.9	0	0	0

Table G.2 Optimisation of MSX minimum peak width setting for 20 minute BSA digest run on HCT. (200/100 fmol)

Minimum peak width (mins)	Number of peaks	Average	SD
0.1	772	2.24	0.63
0.2	498	2.20	0.44
0.3	269	2.18	0.41
0.4	134	2.12	0.38
0.5	60	2.08	0.35
0.6	29	1.99	0.35
0.3 peak area	269	2.21	0.39

XCMS Optimisation

Table G.3 Optimisation of XCMS: fwhm setting for 20 minute BSA digest 1 and 2 pmol run on HCT.

	bin			binlin		
Fwhm	No. peaks P<0.05	Average	SD	No. peaks P<0.05	Average	SD
10	79	81.00	247	158	56	8.03
20	123	2.44	1.28	190	2.78	1.30
30	136	2.30	0.75	247	2.46	0.91
40	138	2.31	0.86	237	2.23	0.70
50	141	2.29	0.77	244	2.33	0.65
60	262	2.16	0.79	244	2.33	0.65
80	130	2.21	0.49	237	2.31	0.55
100	221	2.11	0.56	-	-	-
120	108	2.18	0.43	228	2.21	0.45

Table G.4 Optimisation of XCMS: fwhm setting for 20 minute BSA 2 and 4 pmol run on HCT.

	bin			binlin		
Fwhm	**No. peaks P<0.05**	**Average**	**SD**	**No. peaks P<0.05**	**Average**	**SD**
10	83	3.11	1.90	28	7.37	12.4
20	33	5.71	13.40	231	2.78	1.85
30	162	2.49	3.00	277	2.58	1.06
40	179	2.25	0.82	342	2.44	0.84
50	183	2.16	0.54	-	-	-
60	169	2.14	0.51	-	-	-
80	166	2.13	0.47	-	-	-
100	152	2.13	0.49	-	-	-
120	146	2.08	0.44	331	2.12	0.52

Table G.5 Optimisation of of XCMS: fwhm setting for 20 minute BSA 1 and 2 pmol run on FTICR.

	bin			binlin		
Fwhm	**No. peaks P<0.05**	**Average**	**SD**	**No. peaks P<0.05**	**Average**	**SD**
10	267	1.97	0.50	267	1.97	0.49
20	400	1.90	0.39	400	1.90	0.39
30	383	1.86	0.34	94	1.83	0.24
40	363	1.82	0.32	363	1.81	0.32
50	353	1.79	0.33	353	1.79	0.33
60	325	1.77	0.24	325	1.76	0.24
80	302	1.72	0.24	294	1.72	0.21
100	261	1.70	0.22	261	1.69	0.22
120	242	1.66	0.18	242	1.66	0.18

Table G.6 Optimisation of XCMS: fwhm setting for 20 minute BSA 100 and 200 fmol run on HCT.

	bin			binlin		
Fwhm	**No. peaks P<0.05**	**Average**	**SD**	**No. peaks P<0.05**	**Average**	**SD**
10	14	2.04	0.46	14	2.04	0.46
20	28	2.04	0.48	28	2.04	0.47
40	30	1.92	0.58	30	1.91	0.58

I. Kinase activity profiling in *Drosophila*

This is an ongoing collaboration is with the Perrimon lab in the Genetics department at HMS and involves investigating the dynamics of early embryonic phosphorylation-dependent signalling in *Drosophila melanogaster*. The Perrimon lab has developed a method to knock down genes in embryos by embedding mRNA targeting sequences within a microRNA. Previous transgenic *Drosophila* lines used for knock down gene studies were inefficient in the germLine because the long dsRNA hairpins relied on processing by the siRNA pathway. This pathway appears non-functional in the *Drosophila* germLine and the Perrimon lab has circumvented this issue by exploiting the microRNA pathway. Dr. Richelle Sopko in the Perrimon lab uses these short hairpin microRNA (shmiRNA) lines to map phosphorylation-dependent signalling networks. By using well-characterised phospho-specific antibodies to assess the altered distribution and intensity of phospho-proteins in embryos where kinases and phosphatises have been knocked down with shmiRNA, the contribution of the kinome and 'phosphatosome' to early embryonic signalling pathways that trigger these particular phosphorylation events has begun to be assessed. In a similar and more quantitative manner, KAYAK performed on shmiRNA embryos lacking a particular kinase or phosphatase should enable a more refined assignation of a specific kinase or phosphatase to a specific pathway. Following identification of the responsible kinase and phosphatase, epistasis analysis with other pathway component mutants will be used to determine the hierarchy of phospho-relay.

Of the K60, only 18 shared more than 20 % sequence identity with the corresponding *Drosophila* homolog. The overall plan in the future is to determine a kinase activity profile using KAYAK for each stage of development (0 to 24 hours in 2 hours increments) for each mutant kinase. This type of analysis has already been carried out examining mRNA (Figure I.1). However, only the kinase expression based on mRNA is known and there is no indication of the kinase activity. Performing the same experiment using KAYAK would involve a very large number of reactions if all protein kinases (~350) were considered, and currently this would not be practical in terms of instrument time. Particular kinases of interest in specific temporal windows, such as 0-2hrs, have been selected when the proteomic complement is simpler. Only 100-150 of the approximately 300 kinases in the fly genome are present at 0-2 hours embryogenesis.

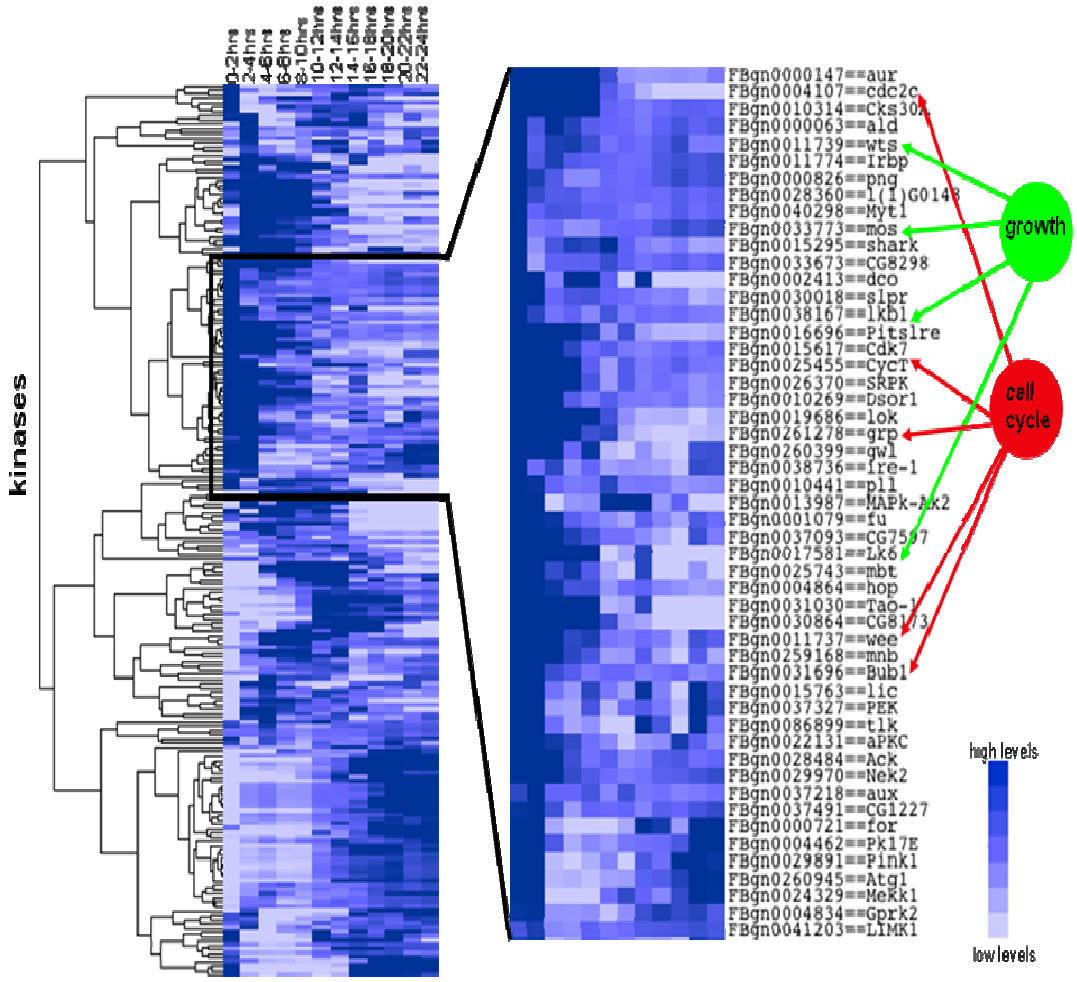


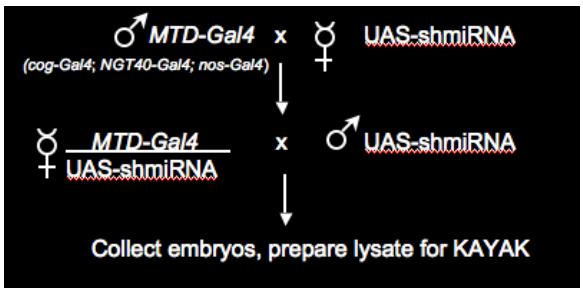
Figure I.1 Kinase transcripts present during *Drosophila* embryogenesis, as assessed by RNASeq by the modENCODE consortium[22].

Table I.1 Kinase and phosphatase genes in *Drosophila* and their human orthologues. These genes were knocked down in the *Drosophila* embryos.

gene		**function**	**human ortholog**	**pathway**
csw	corkscrew	ppase	SHP2	MAPK
dsor	downstream of raf	kinase	MEK1/2	MAPK
hpo	hippo	kinase	MST1/2	Hippo
torso	torso	kinase	RTK but no good orthologue	MAPK
w	white	neg control		
wts	warts	kinase	LATS1/2	Hippo
yki	yorkie	TF	YAP/TAZ	Hippo
egfr	epidermal growth factor receptor	kinase	EGFR	MAPK
pten	pten	ppase	PTEN	PI3K/Akt
Tor	Target of rapamycin	kinase	TOR	PI3K/Akt
S6K	p70 S6 kinase	kinase	S6K	PI3K/Akt
w118	another allele of white	neg control		
sos	suppressor of sevenless	Ras guanyl-nucleotide exchange factor	SOS1/2	RTK
src42A	Src oncogene at 42A	kinase	SRC	RTK
cnk	connector enhancer of kinase suppressor of Ras		CNKSR1/2/3	RTK
dos	daughter of sevenless	SH2 domain binding	no orthologue	RTK
drk	downstream of receptor kinase	SH3/SH2 adaptor	GRB2	RTK
gap1	GTPase-activating protein 1	Ras GTPase activator	RASA2/3	RTK
ksr1	kinase suppressor of ras		KSR1	RTK
shc	SHC-adaptor protein		SHC1/2/3/4	RTK
InR	insulin-like receptor	kinase	IGF1R	RTK
phl	polehole	kinase	Raf	MAPK
ras85D	Ras oncogene at 85D	GTPase	Kras/Hras/Nras	MAPK
rl	rolled	kinase	ERK1/2	MAPK
lacZ	betagalactosidase	negative control		

I.1 Experimental

I.1.1 Drosophila crosses



Virgin female UAS-shmiRNA flies were crossed to male MTD-Gal4[23] flies. The transgenic MTD-Gal4 flies are homozygous for P(Gal4-nos.NGT)40, P(COG-GAL4:VP16) and P(nos-Gal4-VP16) on three different chromosomes. These P-element insertions enable expression of Gal4 (a yeast transcription factor) under the regulation of *NGT*, *cog*, and *nanos* transcriptional elements. These three genes are expressed in the germLine.

The germLine of female progeny generated from this cross bear both Gal4 protein (expressed from *NGT*, *cog*, and *nanos* promoters) and one copy of a chromosome bearing mRNA for the gene we desire to knock down (a kinase or phosphatase) targeting sequences embedded within a microRNA (shmiRNA).The shmiRNA is preceeded by a UAS (upstream activating sequence) for Gal4 (derived from yeast). During oogenesis, Gal4 can bind the UAS sequence, invoking transcription of the shmiRNA. Upon folding, the shmiRNA hairpin is directed to the microRNA machinery, resulting in the down-regulation (or knock-down) of the gene the shmiRNA is targeting. The second cross is merely to induce fertilisation and consequent release of eggs from virgin females bearing eggs with the kinase or phosphatase of interest knocked down.

I.1.2 *Drosophila* embryo extract preparation

Adult flies were discarded after 2 hours in molasses agar/yeast paste vials. Embryos laid on agar/yeast paste were treated for 2 minutes with 50% bleach with shaking and delivered to a small mesh basket in bleach; washed with copious amounts of 0.1% Triton X-100, then water; then transferred with a paintbrush to a 1.5 mL siliconised Eppendorf tube containing 1mL cold KAYAK buffer. After the embryos settled, the buffer containing agar bits was removed under vacuum with a 23gauge needle. Embryos were resuspended in 1mL cold KAYAK lysis buffer with a cut P1000 tip and delivered to 1mL glass dounce homogeniser. All buffer was removed using a vacuum/23gauge needle, avoiding embryos. 100ul KAYAK buffer was added to ~20 µL packed embryos. Following 11 strokes of tight pestle, lysate was transferred to a cold 1.5mL Eppendorf tube. The crude extract was spun at 13kprm for 10min in a cold microfuge. The supernatant was removed and transferred to a new Eppendorf tube. 1µL and 2 µL was taken for a Bradford assay and extract frozen in 20µg aliquots of 20µL (topped up with KAYAK buffer) in liquid N_2. Frozen aliquots were stored at -80^{0}C.

I.1.3 *Drosophila* cell culture extract preparation

Kc167 and clone8 *Drosophila* cells were plated at $2x10^6$ cells/mL in a 6 well plate. The following day insulin was added to 25 µg/mL for 10 minutes. Cell culture medium was aspirated andthe cells washed 1x with 1mL PBS. Cells were suspended in 0.5mL cold KAYAK buffer and incubated on ice for 5 minutes. Crude extract was transferred to a 1.5mL Eppendorf tube and spun at 13krpm for 10 min at 4^0C in a cold microfuge.The supernatant was removed and transferred to a new Eppendorf tube. 1 µL and 2 µL was taken for a Bradford assay and extract frozen in 20 µg aliquots of 20 µL (topped up with KAYAK buffer) in liquid N_2. Frozen aliquots were stored at -80^0C.

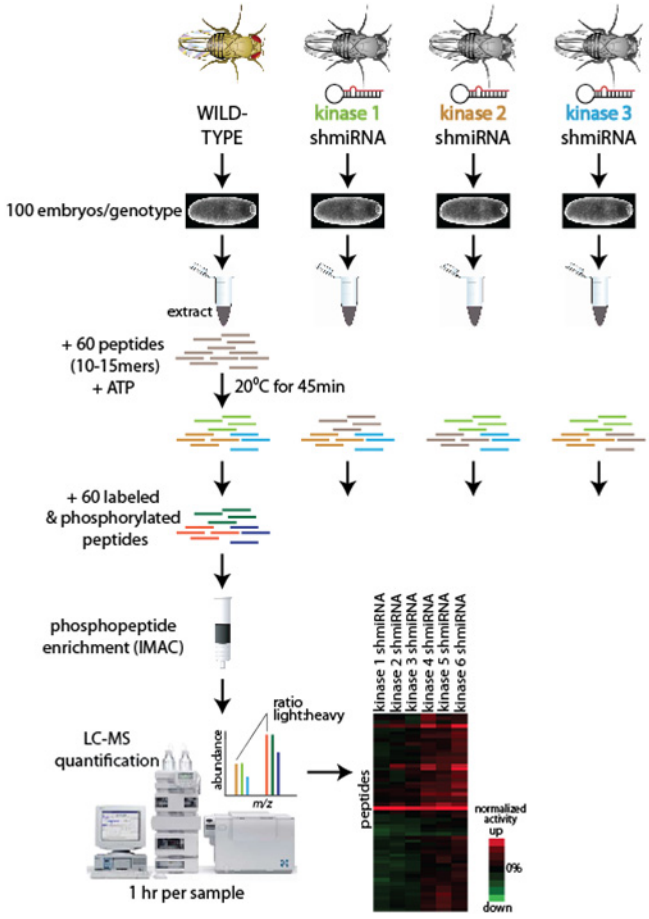


Figure I.2 Workflow schema for KAYAK analysis for the comparison of *Drosophila* wild type and with kinases knocked down.

I.2 Results

KAYAK had not previously been performed on *Drosophila* embryos. Initial 'proof-of-concept' experiments were performed to determine if significant kinase activity was present at different times of embryogenesis and whether significant differences were observed between the WT (wild type) and mutant samples. The samples included: WT embryos (w118) 0-2 hours and 0-4 hours, *torso* shmiRNA mutant embryos (0-2 hours) which had the Torso kinase knocked down, and *corkscrew* shmir mutant embryos (0-2 hours) which had the Corkscrew phosphatase knocked down. Significant kinase activity was observed for many peptides at (53 of the K60) and significant differences can be observed between the mutants (Figure I.3).

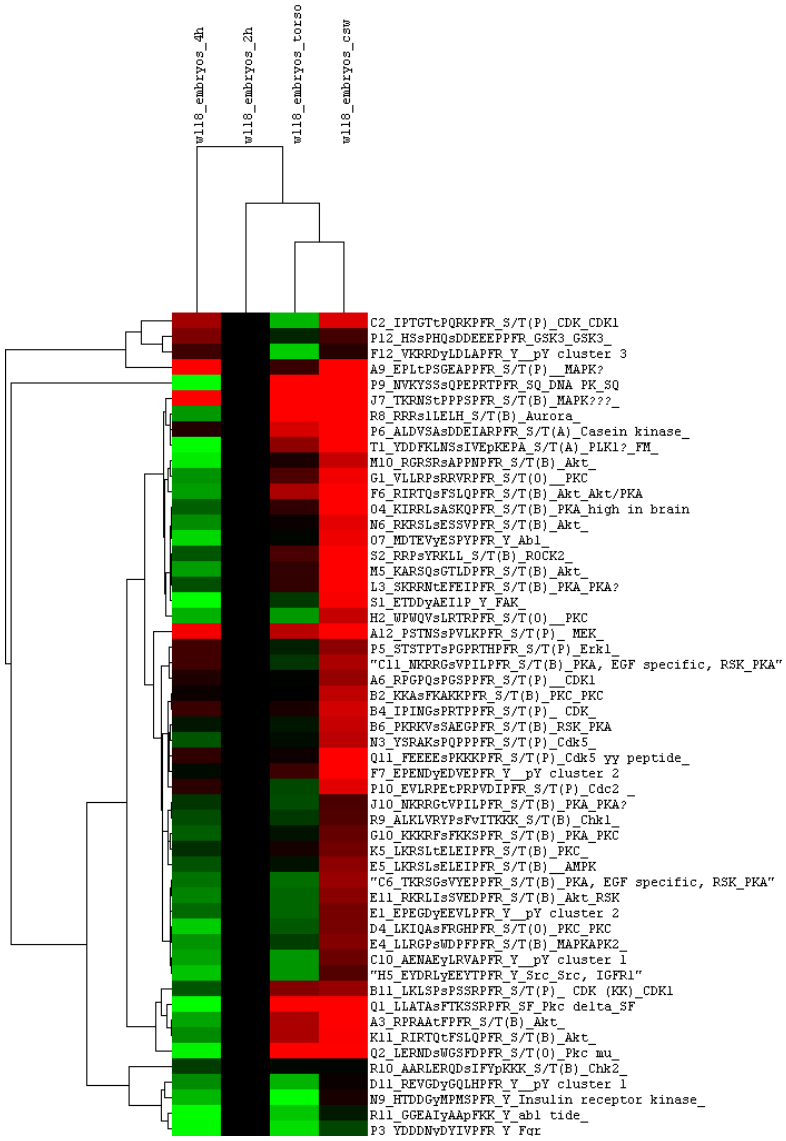


Figure I.3 Profiles of kinase activities using KAYAK with K60 for *Drosophila* embryos 0-4 hours WT, 0-2 hours WT, 0-2 hours *torso* mutant, 0-2 hours *csw* mutant. Kinase activities normalised to w118_2h.

KAYAK was also carried out on *Drosophila* Kc167 cells with and without insulin stimulation. Insulin is known to activate AKT, and increased kinase activity of A3, N6, and K11, which are known AKT

substrates, was observed with the insulin treated cells. Interestingly, AKT activation inhibits GSK3 activity and reduced phosphorylation of P12, a GSK3 substrate peptide from glycogen synthase, was observed, as would be expected (Figure I.4).

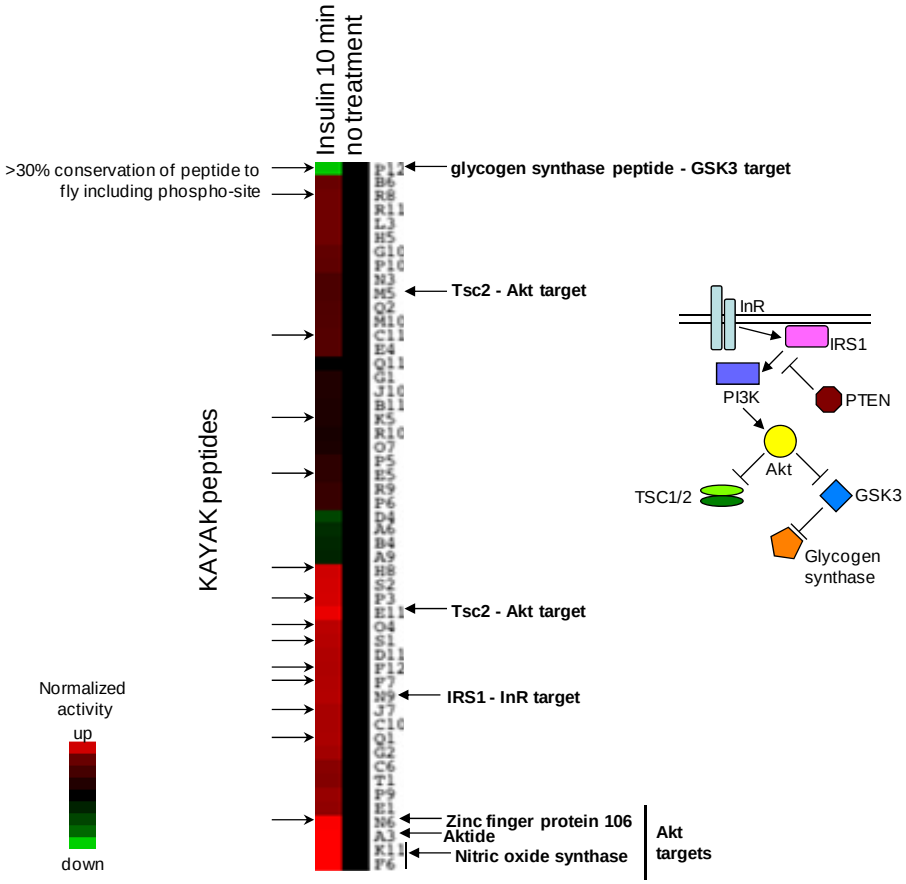


Figure I.4 Profiles of kinase activities using KAYAK with K60 for *Drosophila* cells lines Kc167 with and without 10 minute insulin treatment.

Various kinases and phosphatases were knocked down using shmiR, and the kinase activity was determined using the K60 substrate set (Table 1.1 and Figure I.2). In general, there is a correlation between knocked down genes in the same pathway, as would be expected since these would lead ultimately to similar kinase signatures. The knocked down genes torso, csw and dsor, all from the RTK signalling pathway, cluster together as do warts and hippo from the hippo pathway. Knocked down PTEN increases AKT activity, and indeed increased phosphorylation of the AKT substrates A3 and K11 was observed (Figure I.5).

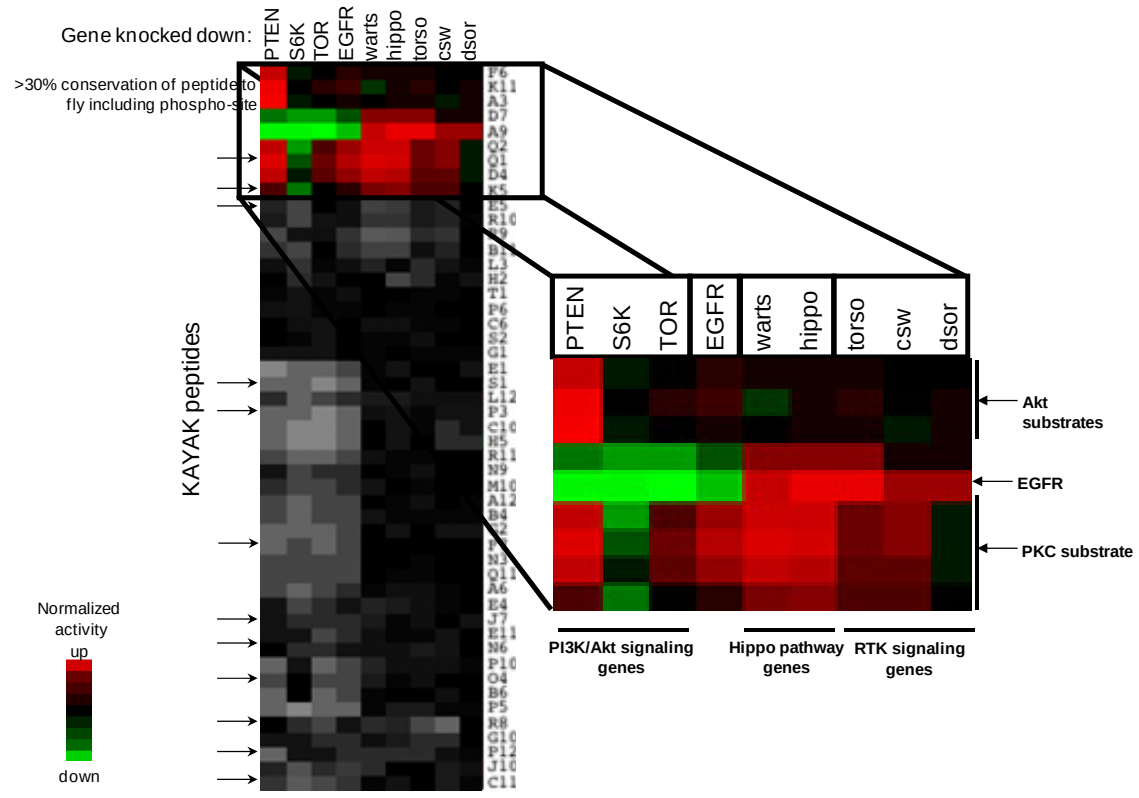


Figure I.5 KAYAK profiling of *Drosophila* shmiR embryo extracts.

The peptide substrates were designed as human target kinases, and yet many of the peptide substrates are phosphorylated by *Drosophila* kinases, so these proof of concept experiments show that KAYAK can be applied to this species to assess the activity of kinases. Many of the kinases and the motifs that they recognise are conserved throughout a wide range of relatively distantly related evolutionary lines, and the use of such peptide target approaches should be able to decipher the motif conservation in a wide range of species.

J. KAYAK in mice treated with fibroblast growth factor (FGF) and mice with targeted disruption of FGF

J.1 Experimental

Mouse livers were provided by Dr. Patty Chui from the Flier laboratory at HMS. **Generation of mice treated with FGF21:** Human FGF21 was expressed in *E. coli* and refolded *in vitro* as described[21]. For analysis of acute signalling events in the liver, recombinant FGF21 was administered via the inferior vena cava to anesthetised adult mice. Briefly, mice were anesthetized via intraperitoneal injection of a ketamine/xylazine cocktail. The peritoneal cavity was then exposed and either FGF21 (100 ng/gram of mouse weight) or saline was injected directly into the inferior vena cava in a total volume of 20 µL. After 10 minutes, the liver was dissected, flash frozen in liquid nitrogen and stored at -80ºC until further analysis. **Generation of mice with targeted disruption of *Fgf21***: FGF21 KO mice were generated by Eli Lilly (Lilly Research Laboratories, Indianapolis, IN). To generate mice with targeted disruption of the *Fgf21* locus, a 6.5 kb *NheI* genomic fragment containing all three exons of the mouse FGF21 gene was subcloned and used as a gene targeting vector by replacing part of exon 1 (30bp downstream of the ATG), all of exon 2, and the 5' region of exon 3 with a neomycin resistance gene (pGTN29; New England Biolabs, Ipswich, MA), thus deleting approximately 1200 bp of the genomic FGF21 sequence. Founder mice were subsequently backcrossed onto the C57Bl/6 line at least 10 times before investigation. **Mouse treatment**: All procedures were approved by the Beth Israel Deaconess Medical Center Institutional Animal Care and Use Committee. FGF21 signalling studies were carried out using male wild-type C57Bl/6 mice obtained from Jackson Laboratory (Bar Harbor, ME). FGF21 , and FGF21 KO mice and their WT littermates were bred at Beth Israel Deaconess Medical Center. All mice were maintained in a temperature-controlled environment at 24ºC on a 12-hour light, 12-hour dark cycle (0600-1800h) with *ad libitum* access to food and water. For the diet-induced obesity signalling studies, mice were fed either regular chow (Harlan Teklad F6 Rodent Diet; Madison, WI) or an obesogenic high-fat high-sucrose diet (D12451, Research Diets, New Brunswick, NJ) for at least 22 weeks. FGF21 KO mice and their WT littermates were fed regular chow as above, then euthanized between 24-32 weeks of age by cervical dislocation; they were then exsanguinated by cardiac puncture, and tissues were rapidly dissected and flash frozen in liquid nitrogen until further analysis. Liver homogenate was prepared for KAYAK as described previously in Section 2.7.2. Two experiments were carried out (Figure J.1). Experiment 1: KAYAK on lean and obese mice with and without FGF21 treatment. KAYAK was performed (in triplicate) on two biological replicates for lean and obese, FGF21 treated and untreated mice, using 20 µg of tissue homogenate for each KAYAK reaction (Figure J.1). KAYAK was performed using the classical 'eppendorf' format as described in Experimental procedures Section 2.6.2.1. Experiment 2: KAYAK on wild type and FGF21 Knock out

mice. KAYAK was performed (in triplicate) on two biological replicates for wild type and FGF21 KO mice using 20 µg of tissue homogenate for each KAYAK reaction (Figure J.1). KAYAK was performed using the classical 'eppendorf' format as described in Experimental procedures Section 2.6.2.1

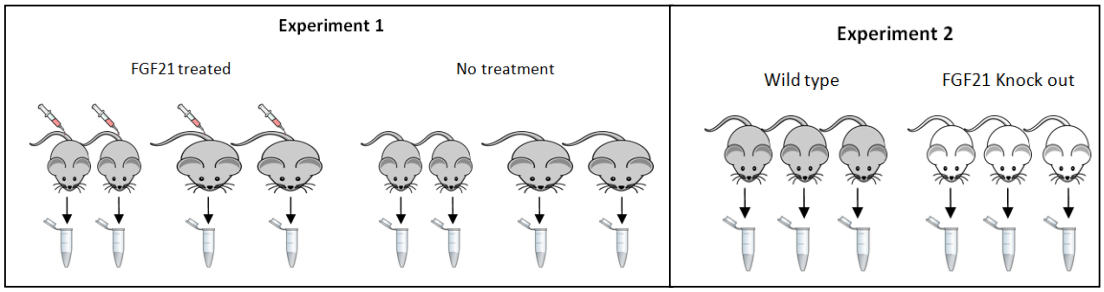
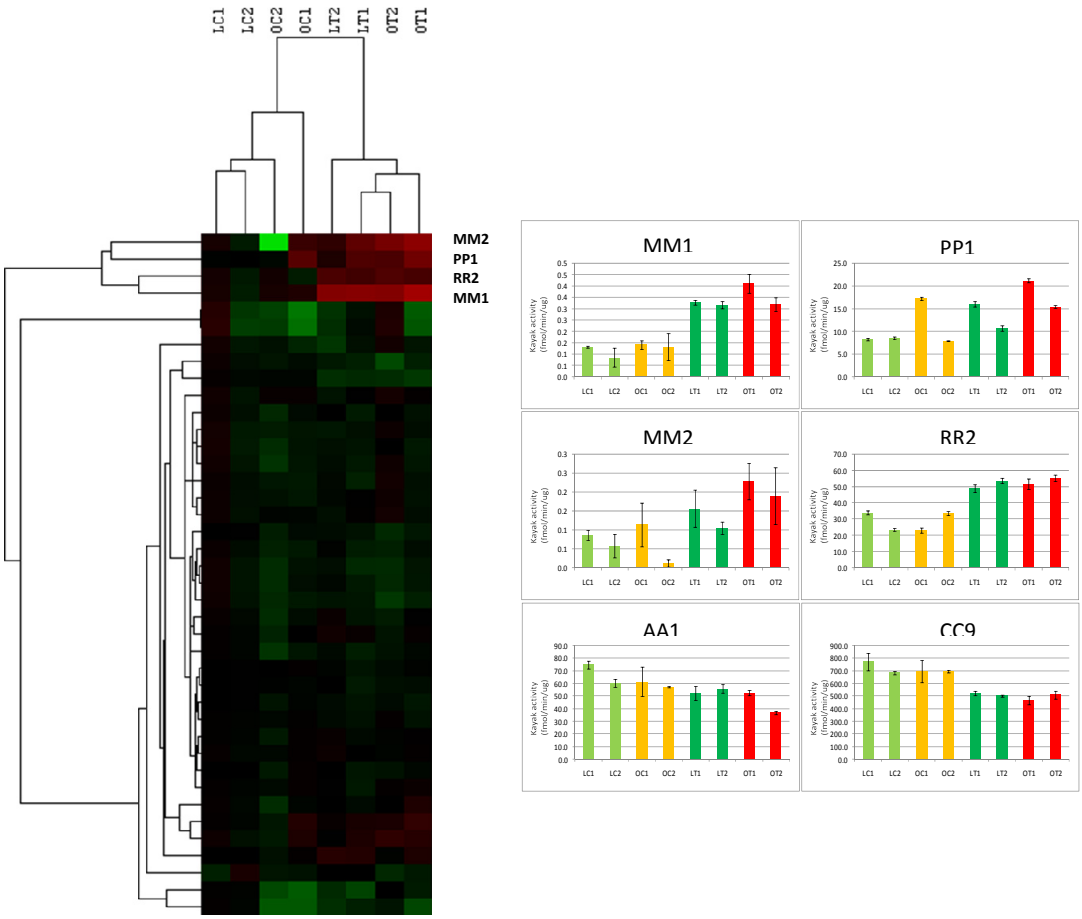


Figure J.1 Schema of mouse treatments with FGF21. Experiment 1 involved a comparison of four groups of mice: lean and obese with and without FGF21 treatment. Experiment 2 involved a comparison of two groups of mice: wild type and FGF21 knock out.

J.2 Kinase activity profiles in mice treated with fibroblast growth factor

Kinase activity profiling was performed on livers from lean and obese, FGF21 (fibroblast growth factor 21) treated and untreated mice. The treated and untreated mice form two clusters based on their kinase activity profiles (Figure J.2). Overall, very few differences in kinase activities are observed. However, increased kinase activity for the peptides MM1, MM2, RR2 and PP1 are observed (Figure J.2). The kinase responsible for the phosphorylation of MM1 is thought to be Erk1 which is one of the downstream targets of FGF21. MM1 is increased 2.8 fold following treatment and has a very low p value of 2.4×10^{-11} (anova, 2 factor with replication). MM2, a MAPK substrate and RR2, a RSK substrate and PP1, a MEK substrate all have significantly greater activity in the FGF treated mouse and these kinases are in the FGF signalling cascade and should be activated with FGF treatment. Despite the inherent biological variability between the mice significant the treated and untreated mice clearly cluster into two groups based on the treatment and that subtle changes can be observed on a kinase activity background that is largely unchanging. A similar kinase profiling experiment was performed on mice with FGF21 knocked out and wild type. Very few differences were observed between the kinase activities and this is was in keeping with the knockout having a similar phenotype to that of the wild type. It is presumed that since FGF21 was a germline knockout, then other compensation from other FGF receptors has occurred.



Peptide	Sequence	Protein name	Predicted kinase	Treated/Control (OT+LT)/(OC+LC)	P value (anova)
MM1	STSTPTsPGPRTHPFR	Ataxin-2-like protein	Erk1	2.80	2.38x10-11
RR2	RKRLIsSVEDPFR	Tuberous sclerosis 2	RSK	1.83	5.96x10-11
PP1	LLRGPsWDPFPFR	Heat shock 27 kDa protein	MAPKAPK2	1.51	9.22x10-4
MM2	EPLtPSGEAPPFR	Epidermal growth factor receptor	MAPK	2.52	0.02
AA1	NKRRGsVPILPFR	Erythrocyte membrane protein band 4.2	PKA	0.78	1.77x10-4
CC9	LKIQAsFRGHPFR	Neurogranin	PKC	0.70	6.88x10-4

Figure J.2 Profile of kinase activities in mouse livers from lean and obese mice treated with and without fibroblast growth factor. (a) Heat map of kinase activities the 8 mice. Analysis by Pearson correlation hierarchical clustering was used to group similar responders together. Each row represents the phosphorylation rate of a particular peptide normalised to the median of kinase activities for the 8 subjects. (b) Examples of kinase activities for particular peptides from (a). Error bars indicate the standard deviation of 3 replicated kinase reactions. (c) Table illustrating the fold increase and p value (anova) between the treated and control group.

J.3 Kinase activity profiles in mice with knocked out FGF

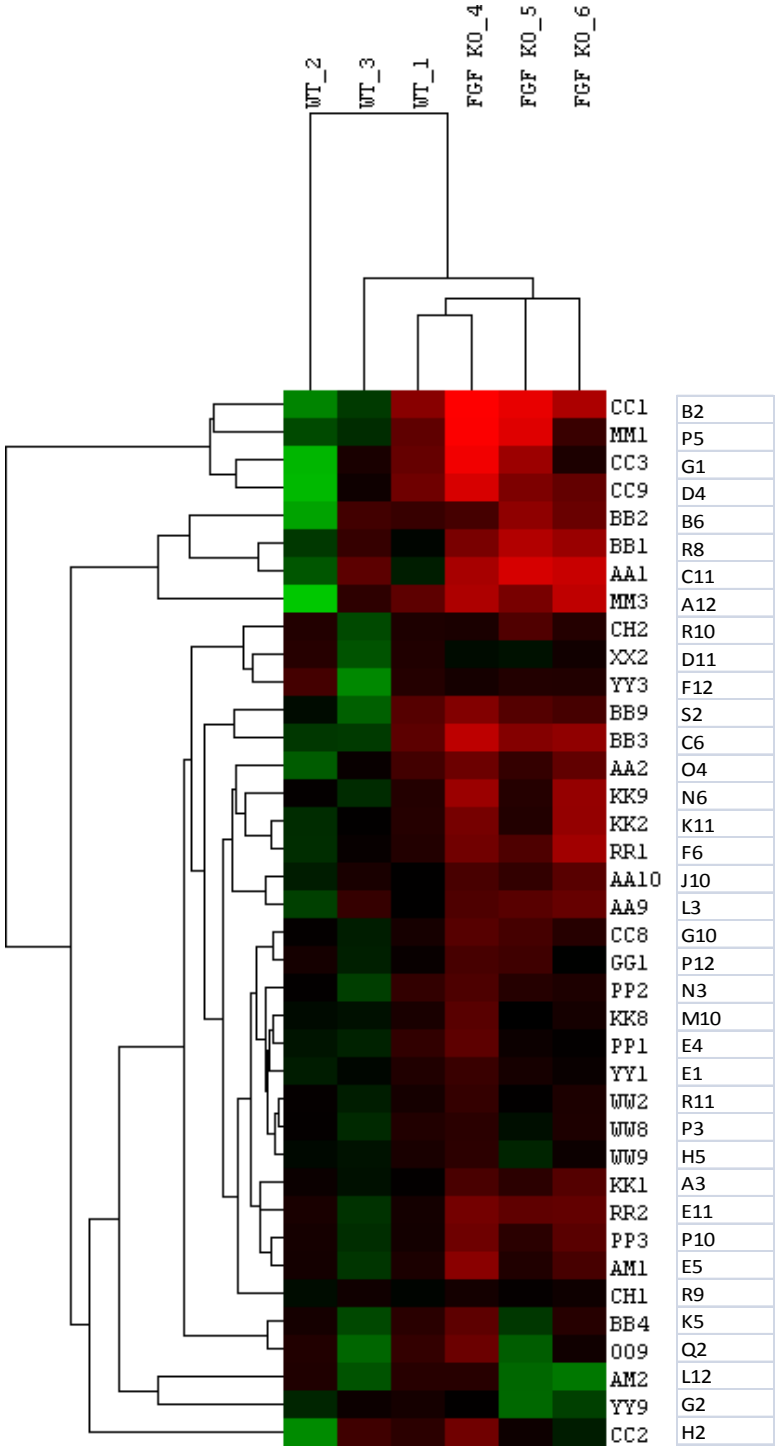


Figure J.3 Profile of kinase activities in mouse livers from control and FGF KO. Heat map of kinase activities. Analysis by Pearson correlation hierarchical clustering was used to group similar responders together. Each row represents the phosphorylation rate of a particular peptide normalised to the median of kinase activities for the 6 subjects.

K. Kinase activity profiling in yeast

KAYAK was performed in yeast (*S. cerevisiae*) and 39 out of the K60 substrate peptides were significantly phosphorylated. None of the tyrosine kinase substrate peptides were phosphorylated, as would be expected since there are no conventional tyrosine kinases in yeast.

KAYAK was performed on asynchronous and nocodazole arrested yeast lysate (Figure K.1). Many differences were observed between the two states. The two MAPK substrate peptides (MM1 and MM3) displayed higher phosphorylation in asynchronous versus nocodazole arrested.

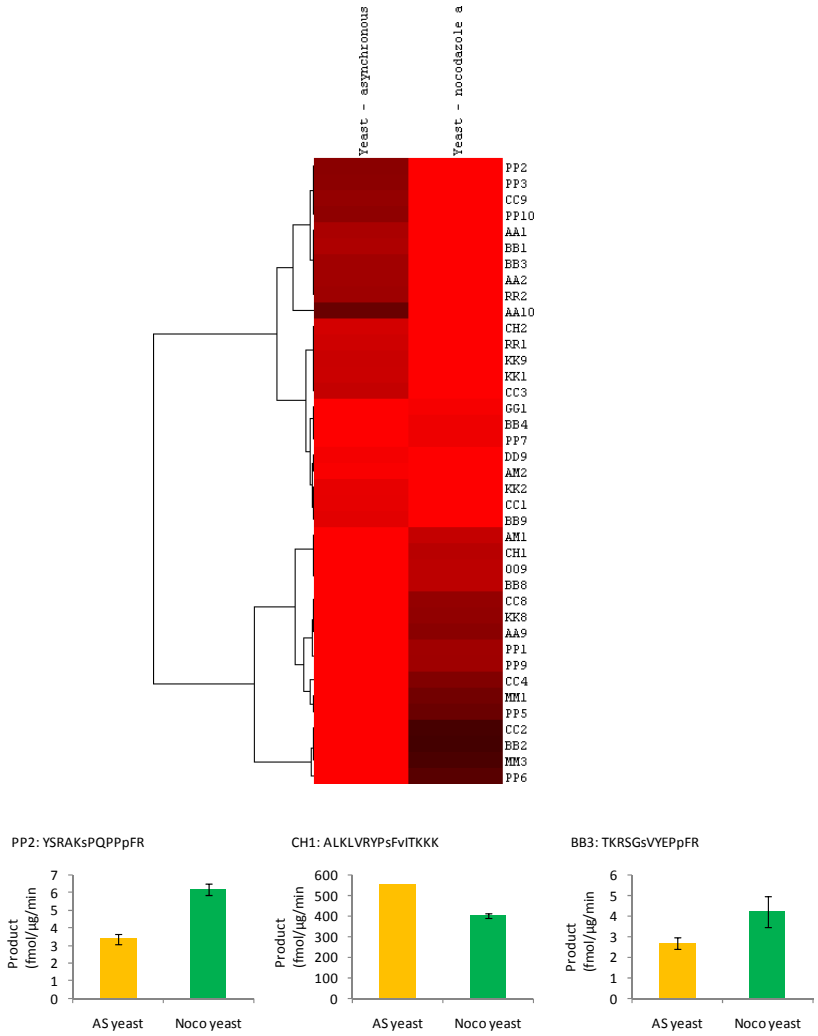


Figure K.1 KAYAK activities in yeast for asynchronous versus nocodazole arrested. 39 out of the K60 substrates showed observable phosphorylation. (a) normalized to the maximum value in row.

L. Correlation of KAYAK with phosphorylation of kinase activation loop

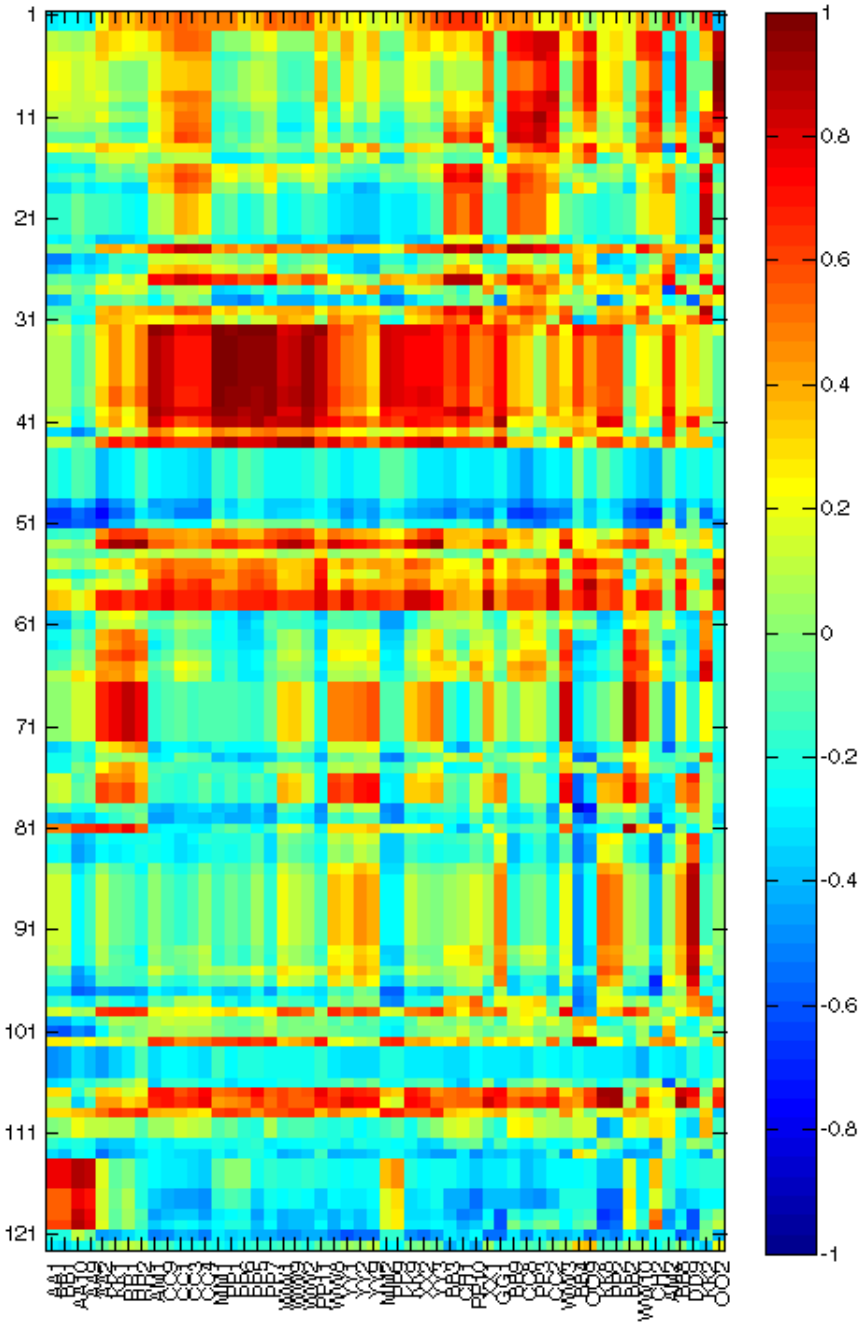


Figure L.1 Correlation profiling using KAYAK-kinase activation loop phosphorylation spectral count. (a) Heat map of Pearson correlation coefficients between the Kinase activity profile (KAYAK) and kinase activation loop phosphorylation spectral count for the 9 mouse tissues. (b and c): Zoomed-in versions of (a).

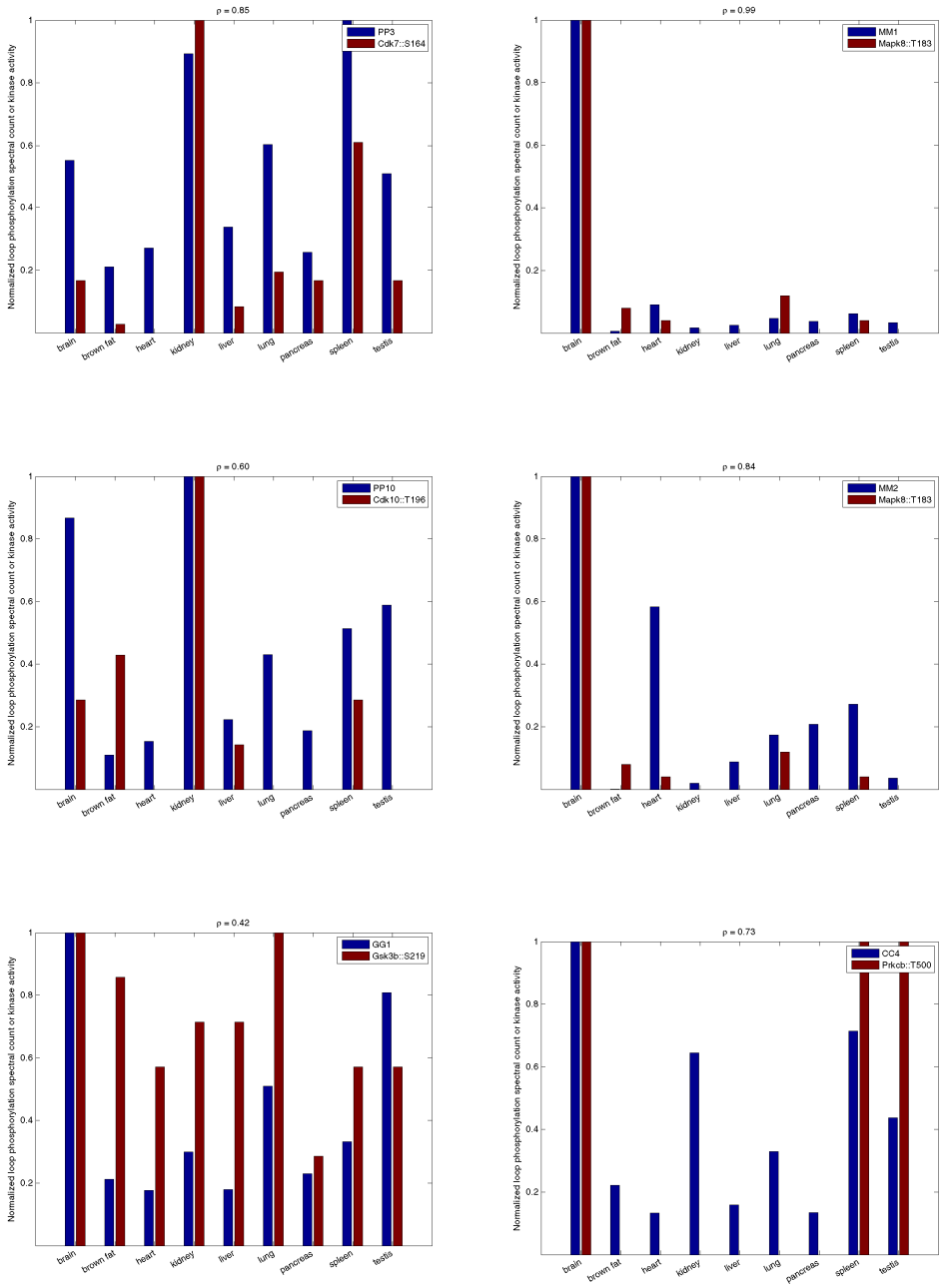


Figure L.2 Correlation profile bar charts of selected known kinase-substrate pairs. The normalised spectral count of the phosphorylated kinase activation loop site and KAYAK activities are plotted for each tissue type. The Pearson correlation coefficient (ρ) for each kinase-substrate pair is given above each plot.

M. Kinase activity correlation profiling

KAYAK correlation profiling of HeLa nocodazole arrested cell lysate

List of proteins for Figure 4.1.1. (too small to see in figure)

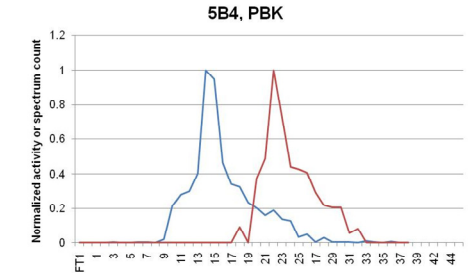
Number	Kinase	Number	Kinase	Number	Kinase
1	MYLK	65	MAP3K5	129	NPR1
2	MYLK2	66	PKMYT1	130	AXL
3	CSNK1A1L	67	STK25	131	BCR
4	EPHB4	68	ROR1	132	PRKD3
5	SCYL2	69	CDK3	133	EIF2AK4
6	MAP2K1	70	EGFR	134	BUB1
7	MAP2K2	71	ERBB4	135	JAK1
8	EPHA5	72	FRAP1	136	SCYL1
9	LYN	73	WNK1	137	SMG1
10	CAMK1D	74	CCNB1	138	STK4
11	STK38	75	DDR2	139	TRIM28
12	PDPK1	76	INSR	140	CDC42BPA
13	MAPKAPK2	77	TP53RK	141	IRAK4
14	FGFR1	78	PRKAA1	142	CDC42BPB
15	CDK4	79	PRKCB1	143	ROR2
16	CDK6	80	CDK7	144	PRKCI
17	MAPK1	81	LIMK2	145	TRIM33
18	MAPK3	82	PKN1	146	GAK
19	RIPK1	83	TRIO	147	GRK5
20	CSK	84	STK24	148	MELK
21	OXSR1	85	RP6-213H19.1	149	TRIM24
22	STK39	86	ROCK1	150	MAP4K4
23	PTK2	87	PKN2	151	ROCK2
24	RIPK2	88	PIK3R4	152	NEK9
25	STK38L	89	WNK3	153	STK3
26	MAP3K3	90	MAPK9	154	NRBP1
27	MERTK	91	TAOK1	155	BUB1B
28	PDGFRB	92	CDC2	156	GRK4
29	CHEK2	93	CDK2	157	CIT
30	STK10	94	CDK5	158	TTBK2
31	ILK	95	GSK3A	159	HSPB8
32	RPS6KA3	96	GSK3B	160	IKBKB
33	MAPKAPK3	97	MARK4	161	TNNI3K
34	FYN	98	MLKL	162	BMP2K
35	MAPK15	99	NEK7	163	RIOK3
36	ACVR2A	100	RPS6KA1	164	RIOK2
37	GUCY2D	101	MAP2K6	165	SRPK1
38	CABC1	102	PTK2B	166	CCNT1
39	MATK	103	PCTK1	167	CDK9
40	OBSCN	104	SRPK3	168	CDC2L2
41	RPS6KA6	105	TTN	169	CLK3
42	YES1	106	STK31	170	MAP4K1
43	AKT2	107	CSNK1A1	171	SRPK2
44	PRKCD	108	PRKACB	172	EPHB1
45	PRKCG	109	ADCK5	173	FLT3
46	EPHA2	110	ARAF	174	KIAA0999
47	MAP2K4	111	HIPK3	175	ADRBK1
48	TAOK3	112	MAPK14		
49	PRKCA	113	ZAK		
50	EIF2AK2	114	PBK		
51	IGF1R	115	PLK1		
52	PAK2	116	TTK		
53	PAK3	117	MAP2K3		
54	PAK1	118	TGFBR1		
55	PRKACA	119	PRKDC		
56	PRKACG	120	CSNK2A1		
57	RIPK4	121	CSNK2A2		
58	AKT1	122	EEF2K		
59	MINK1	123	IRAK1		
60	CASK	124	ULK1		
61	SLK	125	MASTL		
62	CAMKK1	126	LRRK1		
63	MAST1	127	RPS6KA2		
64	TAF1	128	RIOK1		

N. Validation experiments – kinase activites in the presence and absence of lysate

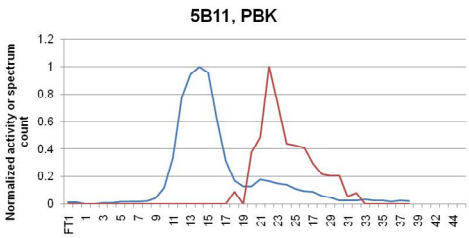
Table N.1 Kinase activities for recombinant kinases (50 ng) in the presence and absence of HeLa nocodazole arrested lysate (20 µg).

	5B4 (fmol)	5B10 (fmol)
EIF2AK2 (50 ng) alone	44 ± 18	0
Lysate (20 µg)	170 ± 60	200 ± 17
EIF2AK2 (50 ng) + lysate (20 µg)	460 ± 130	340 ± 40

An interesting miscellaneous finding during this experiment was that for peptide 5B4, the combination of recombinant PBK and lysate (nocodazole arrested HeLa) led to a 7-fold increase in kinase activity compared to recombinant PBK or lysate alone. PBK was not one of the putative kinase hits for 5B4, which made this an unexpected finding. However, on examining the KAYAK profile, the 5B11 kinase activity had 2 peaks in its profile and the PBK protein profile overlapped with the second peak. A similar effect was also observed for the peptide 5B11, albeit to a lesser extent. There is currently no known explanation for this, but it is a very interesting observation due to the large fold increase in activity and warrants further investigation. Some chemical in the lysate is presumably responsible for activating PBK, but its identity has not been elucidated.



Sample with 5B4	Product formed (fmol)
PBK alone (50 ng)	156 ± 116
Lysate alone (20 ug)	169 ± 58
Lysate (20 ug) + PBK (50 ng)	1065 ± 64



Sample with 5B11	Product formed (fmol)
PBK alone (50 ng)	72 ± 55
Nocodazole alone (20 ug)	328 ± 31
Nocodazole (20 ug) + PBK (50 ng)	504 ± 30

Figure N.1 (A) Kinase activity and protein abundance profiles of PBK and the peptides 5B11 and 5B4. Kinase activities for peptides 5B11 abd 5B4 in the presence of recombinant PBK, with and without lysate.

O. siRNA knockdown of kinases

KAYAK K8 and K60 were performed on the lysates with siRNA knocked down kinases (Figure O.1). The knock down of TP53RK and TOPK does not appear to have any significant effect on either the K8 or K60 peptide substrate set. For KAYAK K60 on the knocked down lysates, decreased phoshphorylation of the peptide substrates for the knocked down kinases was not clear. For MAPK14, decreased phosphorylation of PP1, AM1 and CH2 was observed in the knocked down lysate and these peptides displayed increased phosphorylation with MAPK14 overexpression (Figure). However, there are very few examples of 'expected outcomes', and the results were difficult to interpret, so it is likely that the 'knock downs' were not efficient. No measure of the knock down efficiency of the lysates has currently been determined so it is difficult to assess how effective the knock down was. The knock down efficiency will be assessed in the future using western blotting or quantitative PCR for the kinases which have no commercially available antibodies. Also, there could have been contamination of the lysates with BSA from the cell media. This source of error would be exacerbated through the use of very small lysis volumes from the small culture dishes (12 well). Any media (which contains BSA) not completely removed may therefore 'contaminate' the protein assay which would not correctly measure the amount of lysate present.

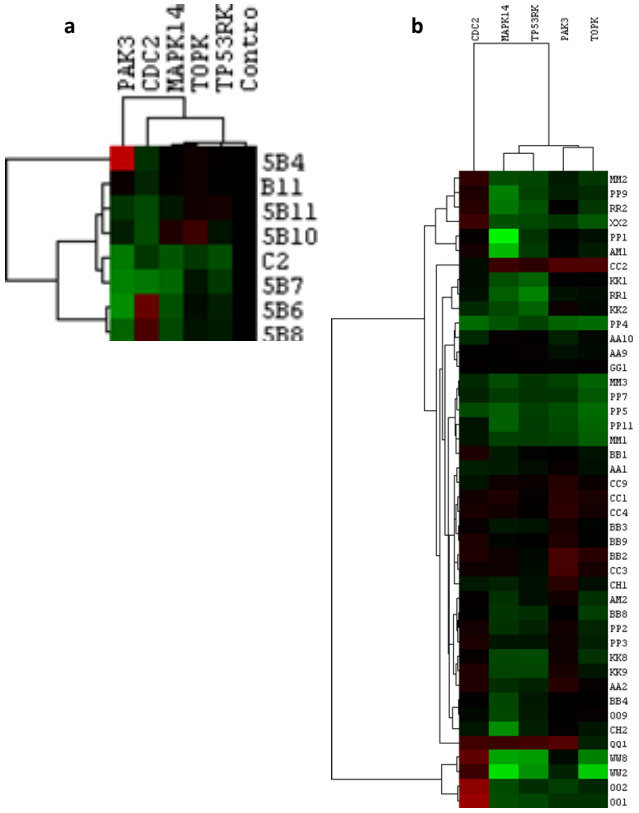


Figure O.1 Heat map of kinase activities for K60 substrates for HEK 293T cells with five siRNA knocked down kinases. Kinase activity is normalised to control cells. Hierarchical clustering grouped similar responders together. Heat map of kinase activities for K8 substrates for HEK 293T cells with five siRNA knocked down kinases. Kinase activity is normalised to control cells. Hierarchical clustering grouped similar responders together.

P. Presence of known interacting partners of PAK3, PLK1 and PBK

P.1 Interactions with PAK3

It is known that CDC42 and RAC stimulate Group I PAK activation. The current model for the activation involves a reversal of a negative regulatory effect by the binding of CDC42 and RAC[24]. In the absence of G protein binding, PAK activity is thought to be repressed by an intramolecular interaction between the regulatory and catalytic domains. On binding by CDC42, RAC and other proteins, the interaction is thought to be disrupted, resulting in an increase in kinase activity[24]. PAK is known to have several binding partners including αPix (Pak-interactive exchange factors), βPix and p50Cool-1.

PAK3/5B11 profiles overlap with part of the RAC3 profile but not with RAC1 or RAC2 which are very low in abundance. There is quite good overlap with CDC2 but the overlap with the other known PAK binding partners is not good (Figure P.1).

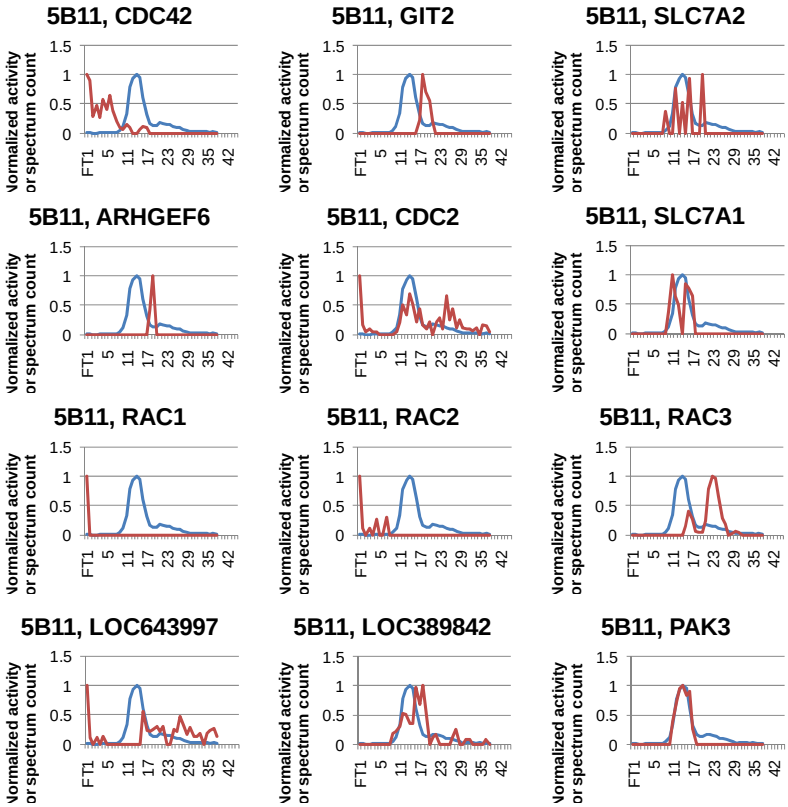


Figure P.1 Correlation profiles of kinase activity for 5B11 and known protein-protein interactions with PAK 3.

P.2 Interactions with PBK

PBK was one of the top hits from the protein correlation profiling for peptides 5B6, 5B7 and 5B8. PBK is reported to form a complex with CDC2/cylin B[25]. However, the protein correlation profiling of PBK does not overlap with CDC2/Cyclin B complex, although there was some overlap between PBK and CDC2 where it is not complexed to CCNB1 (Figure P.2). In siRNA PBK knock down studies, cyclin B1 is reported to be reduced during mitosis whilst CDC2 expression increases slightly[26]. From mutation studies there is evidence that PBK expression and phosphorylation are important for mitotic cyclin B1 expression and it is postulated that PBK helps in recruiting CDC2/cyclin B to mitotic spindles. It is thought that MAPK14 could be a substrate of PBK resulting in MAPK14 activation[26-27] and the protein correlation profiles are very similar, both peaking in fraction 22 (Figure P.2), TP53 (p53) is known to bind to PBK and whilst TP53 itself was not identified, the binding protein TP53BP1 was identified and this fell within the PBK profile as shown in Figure 4.2 Many major substrates of PBK have yet to be found[26].

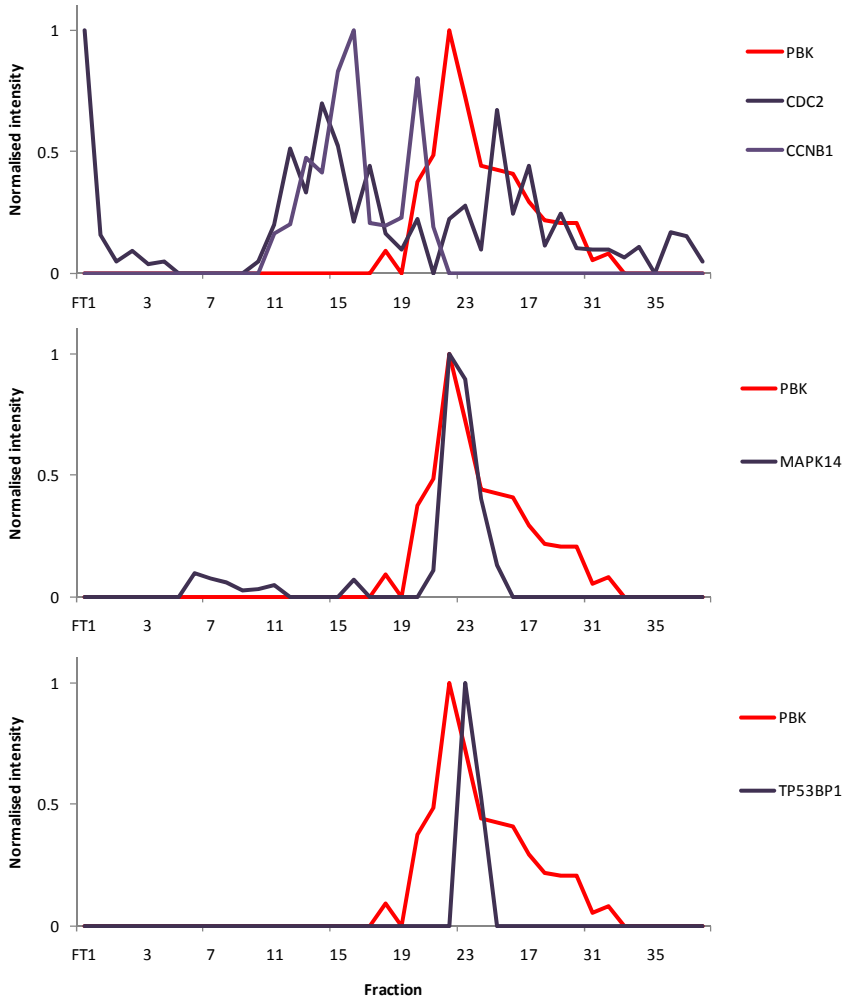


Figure P.2 Protein profiles of (a) PBK, CDC2 and cyclin B, (b) PBK and MAPK14 (c) PBK and TP53BP1.

P.3 Interactions with PLK1

PLK1 associates with various proteins, including MCMs (mini chromosome maintenance proteins)[28] and proteins in the 20S proteasome[29]. There are several overlapping regions in the elution profiles of PLK1 and the various isoforms of the mini chromosome maintenance proteins (MCM) and the proteins of the 20S proteasome (Figure P.3).

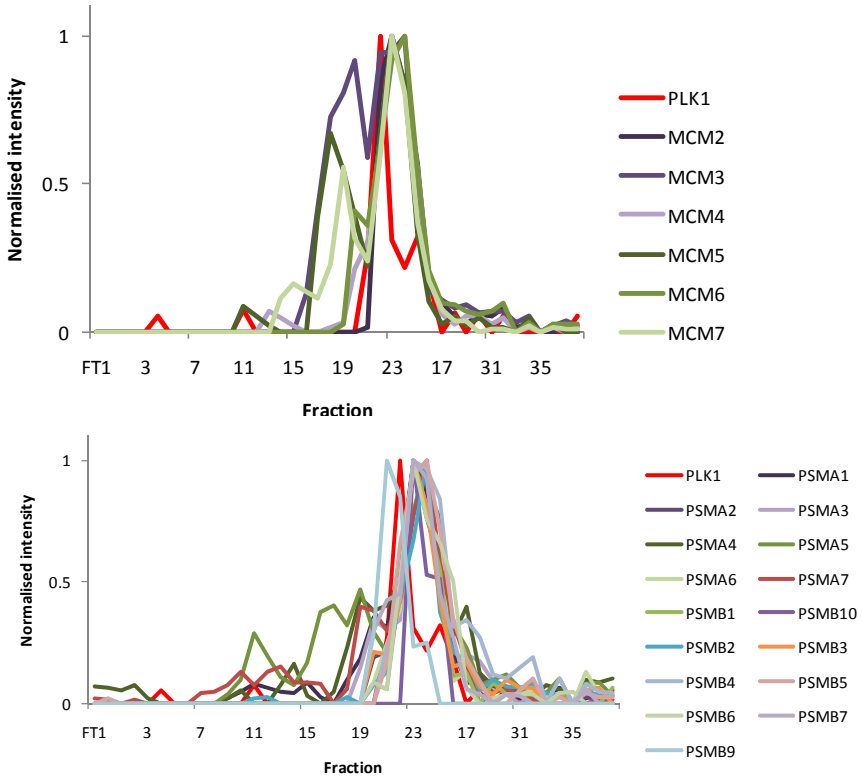


Figure P.3 HeLa cell nocodazole arrested cell lysate separated by anion exchange chromatography into 40 fractions. Elution profiles of (a) PLK1 and various MCM isoforms, (b) PLK1 and 20S proteasome proteins.

Q. Protein sequences of PLK1 substrates

CAMSAP1L1

>sp|Q08AD1|CAMP2_HUMAN Calmodulin-regulated spectrin-associated protein 2 OS=Homo sapiens GN=CAMSAP1L1 PE=1 SV=2
MGDAADPREMRKTFIVPAIKPFDHYDFSRAKIACNLAWLVAKAFGTENVPEELQEPFYTD
QYDQEHIKPPVVNLLLSAELYCRAGSLILKSDAAKPLLGHDAVIQALAQKGLYVTDQEKL
VTERDLHKKPIQMSAHLAMIDTLMMAYTVEMVSIEKVIACAQQYSAFFQATDLPYDIEDA
VMYWINKVNEHLKDIMEQEQKLKEHHTVEAPGGQKSPSKWFWKLVPARYRKEQTLLKQLP
CIPLVENLLKDGTDGCALAALIHFYCPDVVRLEDICLKETMSLADSLYNLQLIQEFCQEY
LNQCCHFTLEDMLYAASSIKSNYLVFMAELFWWFEVVKPSFVQPRVVRPQGAEPVKDMPS
IPVLNAAKRNVLDSSSDFPSSGEGATFTQSHHHLPSRYSRPQAHSSASGGIRRSSSMSYV
DGFIGTWPKEKRSSVHGVSFDISFDKEDSVQRSTPNRGITRSISNEGLTLNNSHVSKHIR
KNLSFKPINGEEEAESIEEELNIDSHSDLKSCVPLNTNELNSNENIHYKLPNGALQNRIL
LDEFGNQIETPSIEEALQIIHDTEKSPHTPQPDQIANGFFLHSQEMSILNSNIKLNQSSP
DNVTDTKGALSPITDNTEVDTGIHVPSEDIPETMDEDSSLRDYTVSLDSDMDDASKFLQD
YDIRTGNTREALSPCPSTVSTKSQPGSSASSSSGVKMTSFAEQKFRKLNHTDGKSSGSSS
QKTTPEGSELNIPHVVAWAQIPEETGLPQGRDTTQLLASEMVHLRMKLEEKRRAIEAQKK
KMEAAFTKQRQKMGRTAFLTVVKKKGDGISPLREEAAGAEDEKVYTDRAKEKESQKTDGQ
RSKSLADIKESMENPQAKWLKSPTTPIDPEKQWNLASPSEETLNEGEILEYTKSIEKLNS
SLHFLQQEMQRLSLQQEMLMQMREQQSWVISPPQPSPQKQIRDFKPSKQAGLSSAIAPFS
SDSPRPTHPSPQSSNRKSASFSVKSQRTPRPNELKITPLNRTLTPPRSVDSLPRLRRFSP
SQVPIQTPSFVCFGDDGEPQLKESKPKEEVKKEELESKGTLEQRGHNPEEKEIKPFESTV
SEVLSLPVTETVCLTPNEDQLNQPTEPPPKPVFPPTAPKNVNLIEVSLSDLKPPEKADVP
VEKYDGESDKEQFDDDQKVCCGFFFKDDQKAENDMAMKRAALLEKRLRREKETQLRKQQL
EAEMEHKKEETRRKTEEERQKKEDERARREFIRQEYMRRKQLKLMEDMDTVIKPRPQVVK
QKKQRPKSIHRDHIESPKTPIKGPPVSSLSLASLNTGDNESVHSGKRTPRSESVEGFLSP
SRCGSRNGEKDWENASTTSSVASGTEYTGPKLYKEPSAKSNKHIIQNALAHCCLAGKVNE
GQKKKILEEMEKSDANNFLILFRDSGCQFRSLYTYCPETEEINKLTGIGPKSITKKMIEG
LYKYNSDRKQFSHIPAKTLSASVDAITIHSHLWQTKRPVTPKKLLPTKA

PB1

>sp|Q86U86|PB1_HUMAN Protein polybromo-1 OS=Homo sapiens GN=PBRM1 PE=1 SV=1
MGSKRRRATSPSSSVSGDFDDGHHSVSTPGPSRKRRRLSNLPTVDPIAVCHELYNTIRDY
KDEQGRLLCELFIRAPKRRNQPDYYEVVSQPIDLMKIQQKLKMEEYDDVNLLTADFQLLF
NNAKSYYKPDSPEYKAACKLWDLYLRTRNEFVQKGEADDEDDDEDGQDNQGTVTEGSSPA
YLKEILEQLLEAIVVATNPSGRLISELFQKLPSKVQYPDYYAIIKEPIDLKTIAQRIQNG
SYKSIHAMAKDIDLLAKNAKTYNEPGSQVFKDANSIKKIFYMKKAEIEHHEMAKSSLRMR
TPSNLAAARLTGPSHSKGSLGEERNPTSKYYRNKRAVQGGRLSAITMALQYGSESEEDAA
LAAARYEEGESEAESITSFMDVSNPFYQLYDTVRSCRNNQGQLIAEPFYHLPSKKKYPDY
YQQIKMPISLQQIRTKLKNQEYETLDHLECDLNLMFENAKRYNVPNSAIYKRVLKLQQVM
QAKKKELARRDDIEDGDSMISSATSDTGSAKRKSKKNIRKQRMKILFNVVLEAREPGSGR
RLCDLFMVKPSKKDYPDYYKIILEPMDLKIIEHNIRNDKYAGEEGMIEDMKLMFRNARHY
NEEGSQVYNDAHILEKLLKEKRKELGPLPDDDDMASPKLKLSRKSGISPKKSKYMTPMQQ
KLNEVYEAVKNYTDKRGRRLSAIFLRLPSRSELPDYYLTIKKPMDMEKIRSHMMANKYQD
IDSMVEDFVMMFNNACTYNEPESLIYKDALVLHKVLLETRRDLEGDEDSHVPNVTLLIQE
LIHNLFVSVMSHQDDEGRCYSDSLAEIPAVDPNFPNKPPLTFDIIRKNVENNRYRRLDLF
QEHMFEVLERARRMNRTDSEIYEDAVELQQFFIKIRDELCKNGEILLSPALSYTTKHLHN
DVEKERKEKLPKEIEEDKLKREEEKREAEKSEDSSGAAGLSGLHRTYSQDCSFKNSMYHV
GDYVYVEPAEANLQPHIVCIERLWEDSAGEKWLYGCWFYRPNETFHLATRKFLEKEVFKS

DYYNKVPVSKILGKCVVMFVKEYFKLCPENFRDEDVFVCESRYSAKTKSFKKIKLWTMPI
SSVRFVPRDVPLPVVRVASVFANADKGDDEKNTDNSEDSRAEDNFNLEKEKEDVPVEMSN
GEPGCHYFEQLHYNDMWLKVGDCVFIKSHGLVRPRVGRIEKVWVRDGAAYFYGPIFIHPE
ETEHEPTKMFYKKEVFLSNLEETCPMTCILGKCAVLSFKDFLSCRPTEIPENDILLCESR
YNESDKQMKKFKGLKRFSLSAKVVDDEIYYFRKPIVPQKEPSPLLEKKIQLLEAKFAELE
GGDDDIEEMGEEDSEVIEPPSLPQLQTPLASELDLMPYTPPQSTPKSAKGSAKKEGSKRK
INMSGYILFSSEMRAVIKAQHPDYSFGELSRLVGTEWRNLETAKKAEYEERAAKVAEQQE
RERAAQQQQPSASPRAGTPVGALMGVVPPPTPMGMLNQQLTPVAGMMGGYPPGLPPLQGP
VDGLVSMGSMQPLHPGGPPPHHLPPGVPGLPGIPPPGVMNQGVAPMVGTPAPGGSPYGQQ
VGVLGPPGQQAPPPYPGPHPAGPPVIQQPTTPMFVAPPPKTQRLLHSEAYLKYIEGLSAE
SNSISKWDQTLAARRRDVHLSKEQESRLPSHWLKSKGAHTTMADALWRLRDLMLRDTLNI
RQAYNLENV

NFKB2

>sp|Q00653|NFKB2_HUMAN Nuclear factor NF-kappa-B p100 subunit
OS=Homo sapiens GN=NFKB2 PE=1 SV=4
MESCYNPGLDGIIEYDDFKLNSSIVEPKEPAPETADGPYLVIVEQPKQRGFRFRYGCEGP
SHGGLPGASSEKGRKTYPTVKICNYEGPAKIEVDLVTHSDPPRAHAHSLVGKQCSELGIC
AVSVGPKDMTAQFNNLGVLHVTKKNMMGTMIQKLQRQRLRSRPQGLTEAEQRELEQEAKE
LKKVMDLSIVRLRFSAFLRASDGSFSLPLKPVISQPIHDSKSPGASNLKISRMDKTAGSV
RGGDEVYLLCDKVQKDDIEVRFYEDDENGWQAFGDFSPTDVHKQYAIVFRTPPYHKMKIE
RPVTVFLQLKRKRGGDVSDSKQFTYYPLVEDKEEVQRKRRKALPTFSQPFGGGSHMGGGS
GGAAGGYGGAGGGGSLGFFPSSLAYSPYQSGAGPMGCYPGGGGGAQMAATVPSRDSGEEA
AEPSAPSRTPQCEPQAPEMLQRAREYNARLFGLAQRSARALLDYGVTADARALLAGQRHL
LTAQDENGDTPLHLAIIHGQTSVIEQIVYVIHHAQDLGVVNLTNHLHQTPLHLAVITGQT
SVVSFLLRVGADPALLDRHGDSAMHLALRAGAGAPELLRALLQSGAPAVPQLLHMPDFEG
LYPVHLAVRARSPECLDLLVDSGAEVEATERQGGRTALHLATEMEELGLVTHLVTKLRAN
VNARTFAGNTPLHLAAGLGYPTLTRLLLKAGADIHAENEEPLCPLPSPPTSDSDSDSEGP
EKDTRSSFRGHTPLDLTCSTKVKTLLLNAAQNTMEPPLTPPSPAGPGLSLGDTALQNLEQ
LLDGPEAQGSWAELAERLGLRSLVDTYRQTTSPSGSLLRSYELAGGDLAGLLEALSDMGL
EEGVRLLRGPETRDKLPSTAEVKEDSAYGSQSVEQEAEKLGPPPEPPGGLCHGHPQPQVH

Figure Q.1 Protein sequences for CAMSAP1L1, PB1 and NFKB2. SSP and STP motifs are highlighted in yellow and SS and ST are highlighted in red. These sites are potential PBD binding sites if the second S/T becomes phosphorylated.

R.Details on induced gingivitis clinical trial

R.1 Subject recruitment and selection

This clinical trial can fulfil its objectives only if appropriate participants are enrolled. Healthy volunteers will be recruited from an existing database of health volunteers within the Eastman Clinical Investigation Centre. Individuals will contacted by telephone to inquire of potential interest. If interested, they will be sent a patient information sheet and advised to contact the centre if they would like to proceed with screening for eligibility. The following eligibility criteria are designed to select subjects for whom protocol treatment is considered appropriate. All relevant medical and non-medical conditions should be taken into consideration when deciding whether this protocol is suitable for a particular subject.

R.1.1 Inclusion Criteria

Subjects must meet all of the following inclusion criteria to be eligible for enrolment into the trial:
1. Males and females ≥ 18 years of age and in good general and oral health.
2. Volunteers must read, sign and receive a copy of the Informed Consent Form after the nature of the study has been fully explained.
3. A minimum of 20 natural teeth with scorable facial and lingual surfaces. Teeth that are grossly carious, extensively restored, orthodontically banded, abutments, exhibiting severe generalized cervical and/or enamel abrasion, or third molars will not be included in the tooth count.
4. Absence of significant oral soft tissue pathology, excluding gingivitis, based on a visual examination. Absence of any object used for lip, tongue, or other form of oral piercing.
5. Absence of moderate/advanced periodontitis based on a clinical examination.
6. Absence of fixed or removable orthodontic appliance or removable partial dentures.

R.1.2 Exclusion Criteria

Subjects presenting with any of the following will not be included in the trial:
1. History of significant adverse effects following use of oral hygiene products such as toothpastes and mouthrinses.
2. History of uncontrolled diabetes or hepatic or renal disease, or other serious medical conditions or transmittable diseases, e.g. cardiovascular disease or AIDS.
3. History of rheumatic fever, heart murmur, mitral valve prolapse or other conditions requiring prophylactic antibiotic coverage prior to invasive dental procedures.
4. Antibiotic, anti-inflammatory or anticoagulant therapy during the month preceding the baseline exam.
5. History of alcohol or drug abuse.
6. Participation in any dental plaque/gingivitis study involving oral care products, concurrently or within the previous 30 days.
7. Self reported pregnancy or lactation (this criterion is due to oral tissue changes related to pregnancy and nursing which can affect interpretation of study results).
8. Other severe acute or chronic medical or psychiatric condition or laboratory abnormality that may increase the risk associated with trial participation or investigational product administration or may interfere with the interpretation of trial results and, in the judgment of the investigator, would make the subject inappropriate for entry into this trial.
9. History of any known food allergies.

R.2 Löe and Silness Gingival Index (GI)

Gingivitis will be assessed at baseline and 2 weeks (Final-Day15) by the Loe and Silness Gingival Index on three surfaces buccal and three surfaces lingual marginal gingivae and interdental papillae of all scorable teeth by the calibrated examiner:

0 - Normal (absence of inflammation).

1 - Mild inflammation (slight change in colour, little change in texture) of any portion of the gingival unit but no bleeding when running a periodontal probe along entrance to the gingival crevice

2 - Moderate inflammation (moderate glazing, redness, edema, and/or hypertrophy) with bleeding of the gingival unit when running a periodontal probe along the entrance to the gingival crevice.

3 - Severe inflammation (marked redness and edema/hypertrophy, or ulceration) with spontaneous bleeding of the gingival unit when running a periodontal probe along the entrance to the gingival crevice.

R.3 Silness-Löe Plaque Index

Plaque area will be scored at baseline and 2 weeks (Final-Day 15) by the calibrated dental examiner, using the Silness-Löe Plaque Index, three buccal surfaces and three lingual surfaces of all scorable teeth following disclosing:

0 - No plaque.

1 – No plaque can be observed by the unaided eye but the plaque is made visible on the point of the probe after it has been moved across the tooth surface at the entrance of the gingival crevice.

2 – The gingival margin is covered with a thin to moderately thick layer of plaque visible to the naked eye.

3 – Heavy accumulation of soft matter, the thickness of which fills out the niche produced by the gingival margin and the tooth surface. The interdental area is filled with soft debris.

S. T-test of gingivitis *vs* healthy

t-test p values for comparison of saliva from individual healthy and gingivitis subjects. No proteins were found to be significant.

Protein	t-test p-value				
AZGP1	0.04	MCTP1	0.27	TXN	0.60
IGKV3-20	0.06	NAIP	0.27	LYZ	0.64
HP	0.10	OPA1	0.27	HBB	0.66
LCN2	0.12	PGD	0.27	HBA1	0.67
IGJ	0.12	PGLYRP1	0.27	ALB	0.68
LOC124220	0.12	PIWIL1	0.27	SMR3B	0.72
HIST2H2BE	0.14	POLR3GL	0.27	LTF	0.74
AMY1A	0.15	RAD54B	0.27	PFN1	0.74
TKT	0.16	RNF149	0.27	AMY2B	0.75
PIP	0.16	RTN4IP1	0.27	LOC440786	0.76
DSG3	0.17	SERPINB1	0.27	BPIL1	0.77
CRISP3	0.17	SLC27A4	0.27	CSTB	0.77
LYPD3	0.20	SYT8	0.27	IGHG2	0.78
ACTB	0.20	TGOLN2	0.27	CASP3	0.80
RIPK4	0.21	YEATS2	0.27	IGL@	0.80
IGHG1	0.21	ZNF433	0.27	IGHV@	0.81
IGHM	0.23	ZNF774	0.27	GSN	0.82
PIGR	0.25	KLK1	0.28	S100A8	0.85
IGHA2	0.27	AMY1A	0.29	ENO1B	0.86
DEFA1	0.27	F7	0.29	ACTA2	0.86
ADCY1	0.27	TF	0.31	CST2	0.86
ALDOA	0.27	GSTP1	0.32	FABP5	0.89
ARHGDIB	0.27	CST4	0.32	PPIA	0.93
C3	0.27	HBB	0.32	PRTN3	0.93
CA1	0.27	HIST1H4D	0.40	ENO3	0.95
CA6	0.27	IGKV2-40	0.40		
CACNB1	0.27	JMJD3	0.40		
CLU	0.27	KIAA1843	0.40		
CNDP2	0.27	L2HGDH	0.40		
COL11A2	0.27	LCN1	0.40		
CRNN	0.27	LOC729708	0.40		
ENO1	0.27	LPO	0.40		
FGA	0.27	MPO	0.40		
FGB	0.27	NFIL3	0.40		
GOLGB1	0.27	NOL1	0.40		
GPI	0.27	SLPI	0.40		
HPR	0.27	TACC2	0.40		
HSPA8	0.27	TCN1	0.40		
IDH1	0.27	C6orf58	0.45		
IGHG4	0.27	CST1	0.48		
IGKV4-1	0.27	S100A12	0.49		
IGLC2	0.27	S100A9	0.50		
ITGA1	0.27	DKFZp686D0972	0.55		
KIAA0317	0.27	IGKV1-5	0.56		
KLHL34	0.27	GAPDH	0.58		
LCP1	0.27	HSPA1L	0.58		
MCM8	0.27	IGHA1	0.60		

T. Bacterial species identified in induced gingivitis trial

List of the proteins identified in six patient undergoing the induction of gingivitis.

	Number of proteins		Total Spectral count	
	Residue	Supernatant	Residue	Supernatant
Abiotrophia defectiva	2	0	2	0
Acinetobacter baumannii	9	1	111	18
Actinomyces coleocanis	4	0	19	0
Actinomyces naeslundii	1	0	1	0
Actinomyces odontolyticus	22	2	584	14
Actinomyces urogenitalis	8	1	70	1
Agrobacterium radiobacter	2	1	51	12
Agrobacterium tumefaciens	4	0	95	0
Anaerococcus hydrogenalis	1	0	6	0
Atopobium rimae	2	1	31	1
Atopobium vaginae	1	0	1	0
Bacillus amyloliquefaciens	3	0	168	0
Bacillus anthracis	3	0	92	0
Bacillus cereus	1	2	8	10
Bacillus halodurans	1	0	5	0
Bacillus selenitireducens	1	0	2	0
Bacillus sp. B14905	1	0	3	0
Bacillus sp. NRRL B-14911	1	0	25	0
Bacillus sp. SG-1	1	0	18	0
Bacillus stearothermophilus	1	0	1	0
Bacillus subtilis	1	0	1	0
Bacillus thuringiensis	2	0	20	0
Bacteroides caccae	17	0	259	0
Bacteroides capillosus	4	0	17	0
Bacteroides coprocola	5	0	42	0
Bacteroides dorei	1	0	1	0
Bacteroides eggerthii	1	0	2	0
Bacteroides fragilis	16	0	273	0
Bacteroides intestinalis	3	0	21	0
Bacteroides ovatus	3	0	52	0
Bacteroides pectinophilus	3	0	55	0
Bacteroides plebeius	3	0	14	0
Bacteroides stercoris	1	1	42	2
Bacteroides thetaiotaomicron	5	2	62	38
Bacteroides uniformis	4	0	33	0
Bacteroides vulgatus	12	0	187	0
Bartonella bacilliformis	1	1	111	13
Bartonella henselae	1	0	1	0
Bartonella quintana	1	0	61	0
Bdellovibrio bacteriovorus	4	1	123	3
Bifidobacterium adolescentis	2	0	4	0
Bifidobacterium animalis	5	1	98	21
Bifidobacterium catenulatum	1	0	2	0
Bifidobacterium dentium	2	0	2	0
Bifidobacterium longum	7	1	12	3
Bordetella avium	5	1	15	139
Bordetella bronchiseptica	2	1	43	3
Bradyrhizobium japonicum	3	1	22	30
Bradyrhizobium sp.	1	0	340	0
Brevundimonas sp. BAL3	1	0	23	0
Burkholderia ambifaria	4	0	6	0
Burkholderia cenocepacia	6	0	37	0
Burkholderia cepacia	2	0	12	0
Burkholderia mallei	3	0	29	0
Burkholderia phymatum	1	0	120	0
Burkholderia phytofirmans	2	1	14	3
Burkholderia sp.	1	0	1	0
Burkholderia thailandensis	1	0	3	0
Burkholderia xenovorans	2	2	113	119
Campylobacter coli RM2228	1	0	1	0
Campylobacter concisus	4	1	76	3
Campylobacter curvus	2	0	12	0
Campylobacter hominis	1	0	8	0
Catonella morbi	1	0	6	0
Caulobacter crescentus	3	0	56	0
Clostridiales bacterium	3	0	9	0
Corynebacterium aurimucosum	2	0	19	0
Corynebacterium diphtheriae	1	0	3	0
Corynebacterium glutamicum	2	0	4	0
Corynebacterium jeikeium	1	0	1	0
Corynebacterium kroppenstedtii	1	0	1	0
Corynebacterium lipophiloflavum	1	0	9	0
Corynebacterium matruchotii	1	0	7	0
Corynebacterium urealyticum	1	0	3	0
Delftia acidovorans	5	1	58	24
Desulfovibrio desulfuricans	1	0	27	0
Desulfovibrio magneticus	2	1	119	57
Eikenella corrodens	7	0	43	0
Enterobacter sp.	4	0	11	0
Enterococcus faecalis	6	0	37	0
Enterococcus faecium	1	1	3	3
Enterococcus hirae	1	0	10	0
Escherichia coli	3	0	13	0
Escherichia coli O127:H6	1	0	1	0
Escherichia coli O139:H28	9	1	221	6
Escherichia coli O157:H7	3	0	20	0
Eubacterium biforme	3	0	74	0
Eubacterium dolichum	1	0	13	0
Eubacterium eligens	2	0	15	0
Eubacterium hallii	1	1	14	2
Eubacterium rectale	3	0	10	0
Eubacterium siraeum	2	0	16	0
Eubacterium ventriosum	7	1	133	4
Finegoldia magna	1	0	10	0
Fusobacterium nucleatum	10	1	91	9
Granulicatella adiacens	1	0	2	0
Haemophilus ducreyi	14	0	46	0
Haemophilus gallinarum	0	1	0	1
Haemophilus influenzae	32	0	214	0

Haemophilus parainfluenzae	2	0	3	0
Haemophilus parasuis	2	0	3	0
Haemophilus somnus	4	0	12	0
Helicobacter pylori	3	0	6	0
Kingella oralis	7	0	16	0
Klebsiella pneumoniae	1	0	5	0
Kocuria rhizophila	9	0	95	0
Lactobacillus amylophilus	1	0	1	0
Lactobacillus brevis	1	0	3	0
Lactobacillus casei	1	0	46	0
Lactobacillus delbrueckii	2	0	42	0
Lactobacillus gasseri	2	0	51	0
Lactobacillus helveticus	1	1	17	1
Lactobacillus jensenii	1	0	53	0
Lactobacillus plantarum	1	0	2	0
Lactobacillus sakei	1	0	2	0
Lactobacillus salivarius	1	0	12	0
Lactobacillus vaginalis	0	1	0	1
Lactococcus lactis	12	1	158	1
Listeria innocua	1	0	19	0
Listeria welshimeri	1	0	1	0
Mobiluncus curtisii	1	0	4	0
Mobiluncus mulieris	2	0	10	0
Mycobacterium abscessus	2	0	40	0
Mycobacterium avium	1	1	2	1
Mycobacterium bovis	2	1	40	11
Mycobacterium leprae	1	0	1	0
Mycobacterium smegmatis	2	1	64	47
Mycobacterium ulcerans	1	0	10	0
Mycobacterium vanbaalenii	1	0	1	0
Mycoplasma capricolum	2	0	15	0
Mycoplasma mycoides	1	0	1	0
Mycoplasma pulmonis	2	0	15	0
Neisseria flavescens	95	0	382	0
Neisseria gonorrhoeae	94	2	1026	10
Neisseria meningitidis	121	1	917	1
Neisseria perflava	1	0	10	0
Ochrobactrum anthropi	1	0	1	0
Oribacterium sinus	2	1	8	13
Porphyromonas endodontalis	4	1	56	1
Porphyromonas gingivalis	10	1	115	5
Porphyromonas uenonis	6	0	26	0
Prevotella intermedia	1	0	1	0
Prevotella loescheii	1	0	24	0
Propionibacterium acnes	2	0	14	0
Propionibacterium freudenreichii subsp. shermanii	1	0	2	0
Proteus mirabilis	1	0	51	0
Pseudomonas aeruginosa	2	0	22	0
Pseudomonas fluorescens	1	0	29	0
Pseudomonas mendocina	1	0	3	0
Pseudomonas putida	1	0	28	0
Pseudomonas sp. BG33R	1	0	11	0
Pseudomonas stutzeri	2	0	22	0
Ralstonia eutropha	2	1	126	92
Ralstonia metallidurans	2	0	6	0
Ralstonia pickettii	5	0	24	0
Ralstonia solanacearum	7	0	24	0
Rhizobium etli	1	0	2	0
Rhizobium leguminosarum bv. trifolii	1	0	13	0
Rhizobium loti	1	0	2	0
Rhizobium tropici	1	0	5	0
Selenomonas flueggei	16	2	121	42
Selenomonas ruminantium	1	0	53	0
Shuttleworthia satelles	2	2	24	6
Sphingomonas sp. SKA58	1	0	1	0
Staphylococcus aureus	1	1	10	6
Staphylococcus cohnii	1	0	16	0
Staphylococcus epidermidis	1	0	53	0
Staphylococcus haemolyticus	1	0	2	0
Stenotrophomonas maltophilia	1	0	5	0
Streptococcus agalactiae	17	1	296	1
Streptococcus canis	2	1	27	1
Streptococcus equi	11	1	62	2
Streptococcus equinus	5	1	75	13
Streptococcus equisimilis	1	0	75	0
Streptococcus gordonii	15	0	291	0
Streptococcus infantarius	7	0	39	0
Streptococcus intermedius	1	1	48	5
Streptococcus mutans	12	0	135	0
Streptococcus oralis	1	0	4	0
Streptococcus parasanguis	2	0	33	0
Streptococcus parauberis	2	0	8	0
Streptococcus pneumoniae	65	3	780	15
Streptococcus pyogenes serotype	10	1	111	1
Streptococcus salivarius	34	3	199	69
Streptococcus sanguinis	13	4	224	47
Streptococcus suis	5	2	55	12
Streptococcus thermophilus	14	2	119	3
Streptococcus uberis	3	0	19	0
Streptococcus vestibularis	2	0	6	0
Treponema pallidum	1	0	6	0
Treponema phagedenis	1	0	3	0
Treponema saccharophilum	1	0	70	0
Veillonella dispar	21	3	216	29
Xanthomonas campestris	1	0	1	0
Yersinia aldovae	1	0	19	0
Yersinia enterocolitica	3	1	37	7
Yersinia frederiksenii	1	0	3	0
Yersinia pestis	1	0	1	0

U.Appendix references

1. Park, S. et al. Hexameric assembly of the proteasomal ATPases is templated through their C termini. *Nature* **459**, 866-870 (2009).
2. Elias, J.E. Proteome analysis by tandem mass spectrometry: Improvement and biological applications. *PhD dissertation, Harvard University, Cambridge, Massachusetts* (2006).
3. Milner, E., Barnea, E., Beer, I. & Admon, A. The turnover kinetics of major histocompatibility complex peptides of human cancer cells. *Mol Cell Proteomics* **5**, 357-365 (2006).
4. Buchler, N.E., Gerland, U. & Hwa, T. Nonlinear protein degradation and the function of genetic circuits. *Proc Natl Acad Sci U S A* **102**, 9559-9564 (2005).
5. Varshavsky, A. The N-end rule: functions, mysteries, uses. *Proc Natl Acad Sci U S A* **93**, 12142-12149 (1996).
6. Jackson, R.J. Alternative mechanisms of initiating translation of mammalian mRNAs. *Biochem Soc Trans* **33**, 1231-1241 (2005).
7. Kindler, S., Wang, H., Richter, D. & Tiedge, H. RNA transport and local control of translation. *Annu Rev Cell Dev Biol* **21**, 223-245 (2005).
8. Bakalarski, C.E. et al. The impact of Peptide abundance and dynamic range on stable-isotope-based quantitative proteomic analyses. *J. Proteome. Res.* **7**, 4756-4765 (2008).
9. Roelofs, J. et al. Chaperone-mediated pathway of proteasome regulatory particle assembly. *Nature* **459**, 861-865 (2009).
10. Festa, R.A. et al. Prokayrotic ubiquitin-like protein (Pup) proteome of Mycobacterium tuberculosis. *PLoS One* **5**, e8589 (2010).
11. Husi, H., McAllister, F. E., Angelopoulos, N., Butler, V. J., Bailey, K. R., Malone, K., MacKay, L., Taylor, P., Page, A. P., Turner, N. J., Barran, P. E., Walkinshaw, M. Selective Chemical Intervention in the Proteome of Caenorhabditis elegans. *Journal of Proteome Research*, Manuscript accepted (2010).
12. Makarov, A. et al. Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer. *Anal Chem* **78**, 2113-2120 (2006).
13. Perry, R.H., Cooks, R.G. & Noll, R.J. Orbitrap mass spectrometry: instrumentation, ion motion and applications. *Mass Spectrom Rev* **27**, 661-699 (2008).
14. Cooks, R.G., Ouyang, Z., Takats, Z. & Wiseman, J.M. Detection Technologies. Ambient mass spectrometry. *Science.* **311**, 1566-1570 (2006).
15. Scigelova, M. & Makarov, A. Orbitrap mass analyzer--overview and applications in proteomics. *Proteomics* **6 Suppl 2**, 16-21 (2006).
16. Yates, J.R.r., Eng, J.K., McCormack, A.L. & Schieltz, D. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal. Chem.* **67**, 1426-1436 (1995).
17. Schwartz, D. & Gygi, S.P. An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. *Nat. Biotechnol.* **23**, 1391-1398 (2005).
18. Smith, C.A., Want, E.J., O'Maille, G., Abagyan, R. & Siuzdak, G. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Anal. Chem.* **78**, 779-787 (2006).
19. C. A. Smith, E.J.W., G. C. Tong, A. Saghatelian, B. F. Cravatt, R. Abbagyan, G. Siuzdak. in 53rd ASMS Conference on Mass SpectrometrySan Antonio Texas; 2005).
20. Lukas N. Mueller, O.R.A.S.S.L.B.B.M.-Y.B.O.V.R.A.M.M. <B><I>SuperHirn</I></B> - a novel tool for high resolution LC-MS-based peptide/protein profiling. *PROTEOMICS* **7**, 3470-3480 (2007).
21. Kharitonenkov, A. et al. FGF-21 as a novel metabolic regulator. *J Clin Invest* **115**, 1627-1635 (2005).
22. figure, m.c.u. *http://flybase.org/cgi-bin/gbrowse/dmelrnaseq/*.
23. Petrella, L.N., Smith-Leiker, T. & Cooley, L. The Ovhts polyprotein is cleaved to produce fusome and ring canal proteins required for Drosophila oogenesis. *Development* **134**, 703-712 (2007).
24. Bagrodia, S. & Cerione, R.A. Pak to the future. *Trends Cell Biol* **9**, 350-355 (1999).

25. Nandi, A.K., Ford, T., Fleksher, D., Neuman, B. & Rapoport, A.P. Attenuation of DNA damage checkpoint by PBK, a novel mitotic kinase, involves protein-protein interaction with tumor suppressor p53. *Biochem Biophys Res Commun* **358**, 181-188 (2007).
26. Matsumoto, S. et al. Characterization of a MAPKK-like protein kinase TOPK. *Biochem Biophys Res Commun* **325**, 997-1004 (2004).
27. Nandi, A., Tidwell, M., Karp, J. & Rapoport, A.P. Protein expression of PDZ-binding kinase is up-regulated in hematologic malignancies and strongly down-regulated during terminal differentiation of HL-60 leukemic cells. *Blood Cells Mol Dis* **32**, 240-245 (2004).
28. Tsvetkov, L. & Stern, D.F. Interaction of chromatin-associated Plk1 and Mcm7. *J Biol Chem* **280**, 11943-11947 (2005).
29. Feng, Y., Longo, D.L. & Ferris, D.K. Polo-like kinase interacts with proteasomes and regulates their activity. *Cell Growth Differ* **12**, 29-37 (2001).